MIT/WHOI 2004-10

Massachusetts Institute of Technology
Woods Hole Oceanographic Institution

Joint Program in Oceanography/ Applied Ocean Science and Engineering

DOCTORAL DISSERTATION

Gene Discovery and Expression Profiling in the Toxin-Producing Marine Diatom, *Pseudo-nitzschia multiseries* (Hasle) Hasle

by

Katie Rose Boissonneault

September 2004

MIT/WHOI

2004-10

Gene Discovery and Expression Profiling in the Toxin-Producing Marine Diatom, *Pseudo-nitzschia multiseries* (Hasle) Hasle

by

Katie Rose Boissonneault

Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

and

Woods Hole Oceanographic Institution
Woods Hole, Massachusetts 02543

September 2004

DOCTORAL DISSERTATION

Funding was provided by the Woods Hole Oceanographic Institution Academic Programs Office.

Approved for publication; distribution unlimited.

Approved for Distribution:

John J. Stegeman, Chair
Department of Biology

Paola Malanotte-Rizzoli
MIT Director of Joint Program

John W. Farrington
WHOI Dean of Graduate Studies

Gene Discovery and Expression Profiling in the Toxin-Producing Marine Diatom, *Pseudo-nitzschia multiseries* (Hasle) Hasle

by

Katie Rose Boissonneault
B.Sc., University of Massachusetts Dartmouth (1995)
M.Sc., MIT/WHOI (1999)

submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

and the

Woods Hole Oceanographic Institution

September 2004

Signature of Author Katie Rose Boissonneault
Joint Program in Biological Oceanography
Massachusetts Institute of Technology and Woods Hole Oceanographic Institution
August 10, 2004

Certified by David E. Housman
Dr. David E. Housman
Ludwig Professor of Biology
Massachusetts Institute of Technology
Thesis Advisor

Accepted by Mark Ettal for
Dr. John Waterbury
Chair, Joint Committee for Biological Oceanography
Woods Hole Oceanographic Institution

Table of Contents

Abstract		**4**
Acknowledgements		**6**
Chapter 1	**Introduction**	**9**
Chapter II	**cDNA library and EST Database**	**23**
Chapter III	**Gene Expression Profiling**	**68**
Chapter IV	**Synthesis and Future Work**	**162**
Literature Cited		**169**

Gene Discovery and Expression Profiling in the Toxin-Producing Marine Diatom, *Pseudo-nitzschia multiseries* (Hasle) Hasle

by

Katie Rose Boissonneault

submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Abstract

Toxic algae are a growing concern in the marine environment. One unique marine diatom, *Pseudo-nitzschia multiseries* (Hasle) Hasle, produces the the neurotoxin domoic acid, which is the cause of amnesic shellfish poisoning. The molecular characterization of this organism has been limited to date. Therefore, the focus of this thesis was to identify and initiate characterization of actively expressed genes that control cell growth and physiology in *P. multiseries*, with the specific goal of identifying genes that may play a significant role in toxin production.

The first step in gene discovery was to establish a complementary DNA (cDNA) library and a database of expressed sequence tags (ESTs) for *P. multiseries*. 2552 cDNAs were sequenced, generating a set of 1955 unique contigs, of which 21% demonstrated significant similarity with known protein coding sequences. Among the genes identified by sequence similarity were several involved in photosynthetic pathways, including fucoxanthin-chlorophyll a/c light harvesting protein and a C4-specific pyruvate, orthophosphate dikinase. Several genes that may be involved in domoic acid synthesis were also revealed through sequence similarity, for example, glutamate dehydrogenase and 5-oxo-L-prolinase. In addition, the identification of sequences that appear novel to *Pseudo-nitzschia* may provide insight into unique aspects of *Pseudo-nitzschia* biology, such as toxin production.

Genes whose expression patterns were correlated with toxin production were identified by hybridization to a microarray manufactured from 5376 cDNAs. 121 cDNAs, representing 12 unique cDNA contigs or non-redundant cDNAs, showed significantly increased expression levels in *P. multiseries* cell populations that were actively producing toxin. The up-regulated transcripts included cDNAs with sequence similarity to 3-carboxymuconate cyclase, phosphoenolpyruvate carboxykinase, an amino acid transporter, a small heat shock protein, a long-chain fatty acid Co-A ligase, and an aldo/keto reductase. These results provide a framework for investigating the control of toxin production in *P. multiseries*. These transcripts may also be useful in ecological field studies in which they may serve as signatures of toxin production. Prospects for further application of molecular genetic technology to the understanding of the physiology and ecology of P. multiseries is discussed.

Special Thanks to

Jefferson T. Turner, my undergraduate advisor

and

David E. Housman, my graduate advisor

for their support and encouragement.

Acknowledgements: I have been blessed with the support of family, friends, colleagues, and mentors. I thank God for the gift of the people that have been a part of my life during the past three years as I have worked on this research project:

David E. Housman, has encouraged me throughout the past three years, beyond the responsibility expected of an advisor. I am truly thankful for the opportunity and support that David has offered to me. David has been a constant source of enthusiasm and encouragement, beginning nine years ago when he and I began a discussion on the value of applying molecular technology to answer questions in marine ecology. This thesis represents the continuation of that conversation. Thank you, David.

Jefferson T. Turner, my undergraduate advisor, has also been a constant source of support and encouragement throughout my academic career. As an undergraduate, I was enrolled at Umass Amherst as an engineering major. However, I spent my sophmore year at UMass Dartmouth, taking biology classes. During this time, I began working in Professor Turner's lab and going out on research cruises with his field crew. Professor Turner's enthusiasm for both laboratory and field work was contagious, and I remained at UMass Dartmouth to finish my undergraduate degree. "Professor Turner" has respectfully transitioned to "Jeff" during my tenure as a graduate student. Throughout his role as an advisor, Jeff has been open and free with his advice and support, and continues to offer support, today. Thank you, Jeff.

Stephen S. Bates' expertise in *Pseudo-nitzschia* biology was essential to this project and helped guide the research. Steve graciously provided algal cultures, DA analyses, and advice. Steve encouraged me with his friendly, generous approach. Stories of his sons and their heartfelt drive to offer help and love in this world was inspiring and comforting throughout my thesis. Thank you, Steve.

Claude Leger, a member of Steve's lab, kindly offered his assistance by running DA analyses. Thank you, Claude.

Mark Hahn supported me throughout both my master's and doctoral degrees. Mark's practical suggestions have guided me through my research, the writing of my thesis, and through the administrative realities. I truly appreciate Mark's integrity and kindness. Thank you, Mark.

Donald M. Anderson offered a perspective that was invaluable for both my master's and doctoral theses. I am thankful for the insights that Don has shared with me. I hope that as I move forward in my career, I will have the opportunity to continue to learn from Don's experience and knowledge. Thank you, Don.

Dave Kulis, a member of Don's lab, graciously offered his assistance, providing culturing facilities and back-up stock culture maintanence. Thank you, Dave.

Senjie Lin's expertise in molecular ecology has helped to guide my research. Senjie's enthusiasm for my research has been helpful and encouraging throughout my thesis. Thank you, Senjie, for your time and advice.

John Waterbury offered patience, guidance and support in completing my defense and written thesis. Thank you, John.

Judy McDowell, John Farrington, Julia Westwater, and *Marsha Gomes* provided administrative support. Thank you for your support and patience.

Jerry Pelletier welcomed me into his lab and guided me through the construction of the cDNA library. Jerry also provided sequencing facilities for the EST database. I value Jerry's guidance and friendship. Thank you, Jerry.

Isabelle Harvey and *Nhi Nguyen,* members of Jerry's lab, offered their assistance in both the library construction and sequencing. Thank you, both.

Sean Milton offered his expertise to help design and build the cDNA microarray, and *Shirley Li* provided technical assistance. Thank you, Sean and Shirley.

Penny Chisholm and her lab members shared their lab meetings and culture facilities with me. Thank you.

Duaa Mohammed and *Aaron Aslanian* proof-read drafts of my thesis, and offered support and friendship. Thank you, Duaa and Aaron.

Housman lab members offered support and advice throughout the completion of my thesis research; thank you. I would like to specifically acknowledge and thank the following individuals:

Michele Maxwell offered her time and expertise to guide my early efforts in extracting RNA from *P. multiseries*. Michele especially guided me through trouble-shooting the initial RNA extractions as I became familiar with the techniques.

Junne Kamihara has stood by my side, both literally and figuratively, for the past three years. Junne and I have shared each other's successes and failures, as well as lab work. Among other tasks, Junne has proof-read, taken care of my cultures, and run gels for me. Junne's strong faith and honesty have been a gift to me. It's your turn, now!

Janette Knowlton has been a loyal friend over the past two years. Janette appreciates life's challenges and is always willing to help. Janette offered assistance with data entry and proof-reading, and continually offers encouragement and support.

Hitomi Hutzell offered administrative assistance, support, friendship, and cat-sitting.

Myra Coufal offered laboratory assistance, support and friendship.

Adel Tabchy assisted me in writing Pearl Script to transfer my data into spreadsheets.

J. Michael Andresen provided lab and computer assistance.

Connie Lavoie offered support and assistance with administrative challenges.

I have been also been supported by the constant love and friendship of my ***family*** and ***friends***. Thank you all for your faithful support and prayers. I especially thank my mother and father, *Kathleen Rose* and *Donald Boissonneault*, my grandparents, *Alice* and *George Bellerose*, my sister, *Donna*, and her family *Bill*, *Marie*, *Joey*, and *Carolyn Rose Heffernan*, my sister, *Fay*, and her family *Kevin*, and *Karly Roux*, my brother, *Joseph John Boissoneault*, my aunt, *Janie Bellerose*, and my dear friends, *Laura, Tom, Emily*, and *Anna Difonzo*, *Pam, Mike,* and *Abby Neubert, Judith Ann Gregoire, Karen Lee Hunter,* and *Sister Olga Yaqob.*

**Financial support for this research was provided by the Woods Hole Academic Programs Office.

Chapter I

Introduction

Toxic algae have become a growing concern in the study of the marine environment during the past few decades (Anderson, 1994; Sellner, 2003). *Pseudo-nitzschia multiseries* is a particularly interesting toxin-producing alga, as it represents one of the only known species to produce a phycotoxin within the division Bacillariophyta (Bates, 1998). The present study focused on the molecular characterization of *P. multiseries*, with special interest in both its role as a harmful alga and as a member of the diatom community.

Diatoms: Diatoms (Bacillariophyta) represent an important group of bloom-forming eukaryotic phytoplankton (Mann and Droop, 1996). They play a major role in global carbon cycling and nutrient cycling in the marine environment (Werner, 1977; Field et al., 1998; Mann, 1999). One distinguishing characteristic of diatoms is their intricate siliceous cell walls, or frustules. Due to the uptake and processing of silicon that is required to produce these frustules, diatoms play a key role in the biogeochemical cycling of silicon and are responsible for the production of 240×10^{12} moles of silica per year (Treguer et al., 1995).

Toxin-producing diatoms appear to be limited to twelve species that produce the neurotoxin, domoic acid (DA): *Amphora coffeaformis*, *Pseudo-nitzschia multiseries*, *P. pseudodelicatissima*, *P. calliantha*, *P. australis*, *P. seriata*, *P. fraudulenta*, *P. delicatissima*, *P. turgidula*, *P. multistriata*, *P. pungens* and *Nitzschia navis-varingica* (Bates et al., 1998; Bates, 2000). The existence of non-toxic strains of *P. multiseries*, *P. seriata*, *P. australis*, *P. delicatissima*, *P. calliantha* and *P. pseudodelicatissima*, and of toxic strains of the generally non-toxic *P. pungens*, suggests genetic variability among strains of the *Pseudo-nitzschia* species and differences in regulatory factors controlling DA production.

Molecular characterization of the toxin-producing diatom species has been limited. Ribosomal RNA genes have been characterized for phylogeny and field identification studies (e.g. Hasle 1994, 1995; Scholin, 1994; Lundholm, 2002). Molecular phylogeny utilizing ribosomal RNA has contributed to changes in

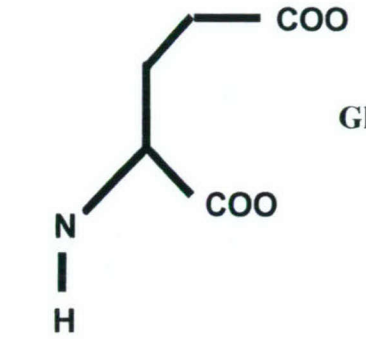
nomenclature of the genus *Pseudo-nitzschia* over the last decade (Hasle, 1994, 1995; Hasle et al., 1996). For example, once considered the same species, *P. multiseries* was distinguished from *P. pungens* due to differences in morphology, physiology and genetic structure of large-subunit ribosomal RNA. Nine DNA microsatellite markers have recently been developed. These markers have been used in field studies to distinguish and analyze relationships among field isolates and in laboratory mating experiments to demonstrate Mendelian inheritance (Evans et al., 2004).

Among the toxin-producing diatoms, the physiology and ecology of *P. multiseries* has been studied the most extensively. Therefore, this species was selected as a model to investigate genes associated with toxin production and overall growth and physiology within this group of marine algae.

Pseudo-nitzschia multiseries: *Pseudo-nitzschia multiseries* is a species of pennate diatom that produces the neurotoxin domoic acid (DA). Production of phycotoxins by diatoms was unknown prior to 1987, when *P. multiseries* first bloomed in Cardigan Bay, Prince Edward Island, Canada (Bates et al., 1989). This initial bloom caused amnesic shellfish poisoning (ASP) in humans who had consumed contaminated blue mussels (*Mytilus edulis*). DA was isolated from extracts of mussels that had been feeding on *P. multiseries* and subsequent studies verified the production of DA by *P. multiseries* and other members of the genus *Pseudo-nitzschia* (Wright et al., 1989). Since 1987, environmental factors influencing DA production in *Pseudo-nitzschia* spp. have been investigated, especially in *P. multiseries* (Bates, Bates et al., 1998). However, the mechanism of DA production and genetic regulation is still not clearly understood.

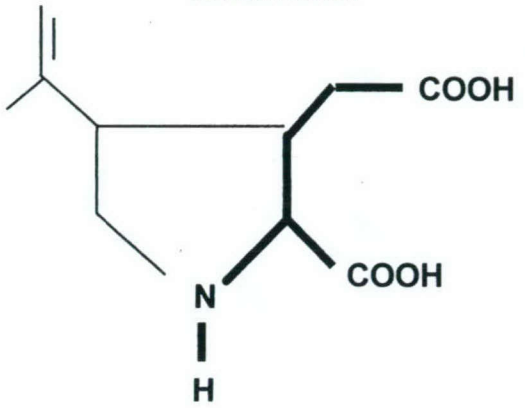
DA is a water-soluble tricarboxylic amino acid with a molecular weight of 311 that includes a proline-like ring containing an isoprenoid and a carboxymethyl side chain (Figure 1-1) (Takemoto and Daigo, 1958). An analog of the neurotransmitters glutamate and kainate, DA predominantly binds to a kainate sub-type of ionotropic glutamate receptor in the central nervous system (Hampson and Manalo, 1998; Berman et al., 2002.) DA has a binding affinity 100 times greater than glutamate and three times

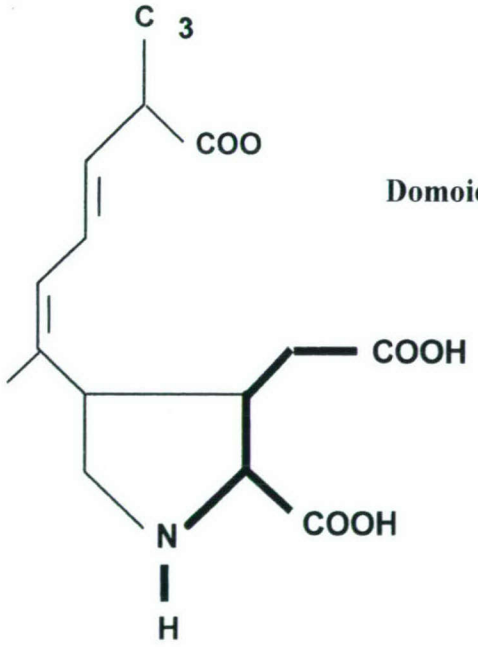
Figure 1-1: Domoic acid and Structural Analogs



Glutamic Acid

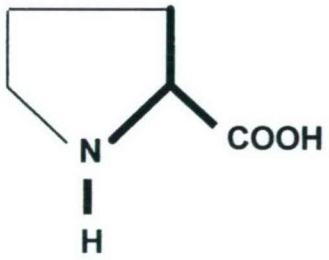
Kainic Acid





Domoic Acid

Proline



greater than kainate. The high affinity binding of DA to glutamate receptors leads to an influx of calcium ions in neurons expressing glutamate receptors, which in turn leads to massive depolarization of these neurons, neuronal swelling, and ultimately cell death (Stewart et al., 1990; Olney, 1994). DA toxicity most severely affects hippocampal nerve cells associated with memory retention, suggesting a functional basis for the memory loss of patients diagnosed with ASP due to DA produced by *P. multiseries* (Todd, 1993).

Characterization of the biosynthetic pathways leading to DA synthesis has been limited. ^{13}C- and ^{14}C- labeling studies suggest a model involving condensation of an activated glutamate derivative from the citric acid cyle with an isoprenoid chain, such as geranyl pyrophosphate, and subsequent cyclization as a possible mechanism to generate DA (Figure 1-2) (Douglas et al., 1992; Ramsey et al., 1998). In separate studies, Smith and colleagues have focused on the relationship of proline to DA metabolism, by measuring amino acid levels to show that proline and DA levels are inversely correlated. They suggest that proline is an upstream precursor to DA, or that DA substitutes for the physiological function of proline. A proposed model showing the hypothesized derivation of 3-hydroxy-glutamate from proline metabolism, which would then lead to DA production, based on the suggestion of Smith et. al. (2001) is shown in Figure 1-3.

Growth Dynamics and DA production: Pseudo-nitzschia multiseries growth rates range from 0.21 to 1.20 divisions per day during the exponential phase in batch culture, while cell yields average between 100,000-300,000 cells/mL, depending on nutrient conditions. Numerous studies on the growth of *P. multiseries* in culture have shown that DA production does not begin until early stationary phase, i.e. toxin is not typically produced during the exponential growth phase (Bates et al., 1989, 1991, 1993, 1995; Subba Rao et al., 1990; Reap, 1991; Douglas and Bates, 1992; Douglas et al., 1993; Lewis et al., 1993). In these studies, cellular DA concentrations reached a peak about one week after the beginning of the stationary phase in batch culture, while the amount of DA released into the culture medium continued to increase throughout the mid- and late- stationary phases (Bates et al., 1991; Pan et al., 1996). In other studies that exposed *P. multiseries* to

Figure 1-2: Proposed pathway for Domoic Acid Biosynthesis, Ramsey et al., 1998

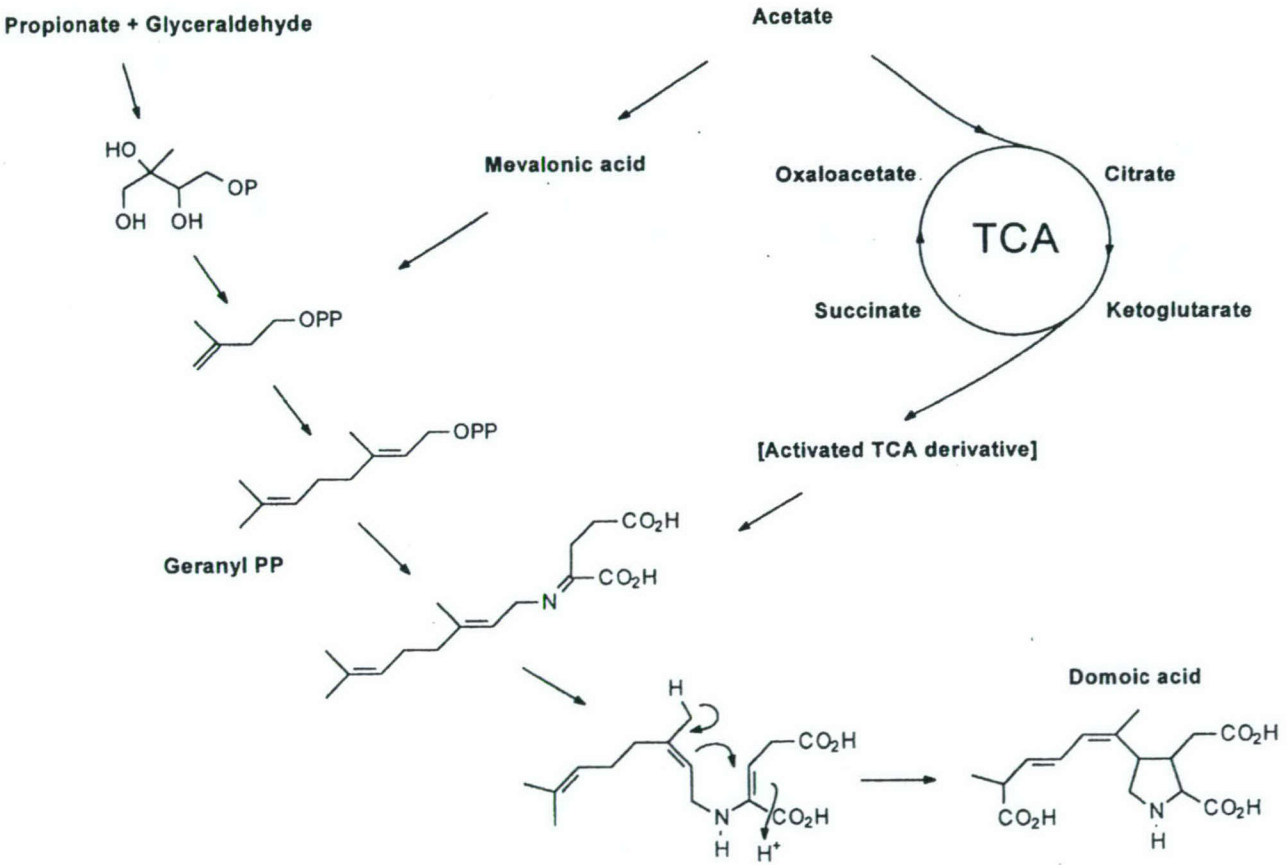
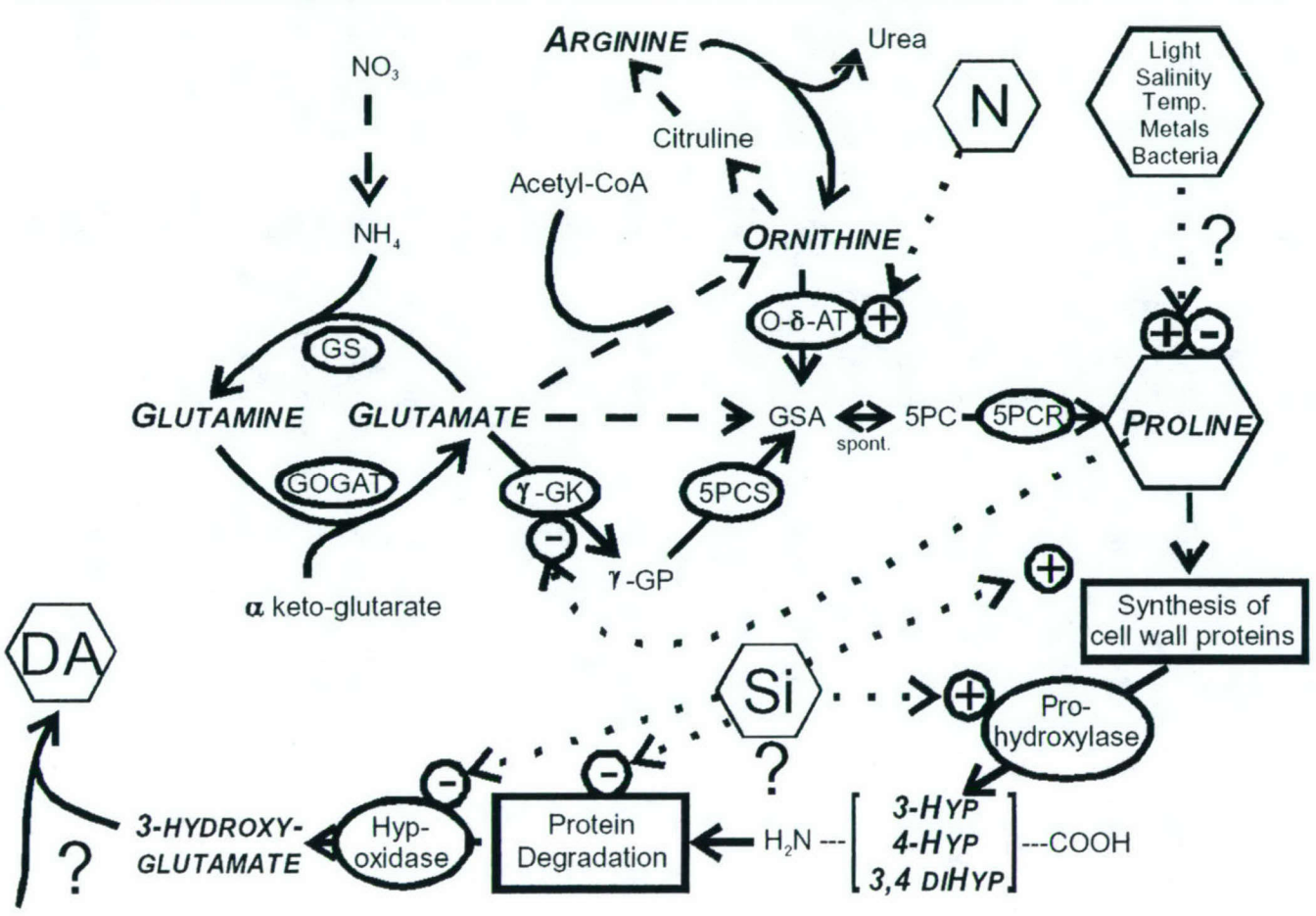


Figure 1-3: Proposed pathway for Domoic Acid Biosynthesis, Smith et al., 2001



conditions in which cell division during mid-exponential phase was slowed relative to normal, cells did produce low levels of toxin (Bates et al., 1993; Pan et al., 1996). Therefore, toxin production appears to be linked to stages in the cell cycle when cell division has stopped or cells are arrested as the division rate of the entire population of cells slows due to some limiting factor (Bates, 1998).

DA production by *P. multiseries* has been associated with physiological stress caused by silicon (Si) limitation. Diatoms require Si for DNA synthesis as well as for frustule construction; Si may therefore become a limiting factor. Bates et al. (1991) and Pan et al. (1996) both showed that the production of DA by *P. multiseries* was inversely correlated with ambient silicate concentration and that DA accumulated in cells when the division rate declined due to depletion of Si. Brzezinski et al. (1990) have shown that Si limitation in diatoms alters the normal progression of cells through the cell cycle (G1, S, G2, M) by arresting cells at the G1/S boundary and in the G2 or M phases. DA production in *P. multiseries* appears to begin at the end of G1 or during the G2 phase of the cell cycle, which correlates with cell cycle arrest due to Si limitation (Pan et al., 1996; Bates and Richard, 1996). Si limitation may impede the progression of the cell cycle by interfering with DNA synthesis. In separate studies, Sullivan and Volcani (1973) showed that cessation of DNA synthesis by Si starvation was caused by a decrease in DNA polymerase and thymidilate (TMP)-kinase activity, but not by a lack of energy or precursors. DNA polymerases A and D are only synthesized in the presence of Si, whereas at least 15 other proteins are formed only in the absence of Si. These results suggest that Si levels affect regulation of gene expression in diatoms (Pan et al., 1998).

Phosphorous (P) limitation has also been implicated as a trigger for DA production (Bates et al., 1991; Pan et al., 1996, 1998). Toxin production was induced in batch culture when phosphate supply was low (<1 μM) and alkaline phosphatase activity (an indicator of P-limitation) was high. In addition, synthesis of DA was depressed by the addition of inorganic P, which stimulated cell growth. In contrast to Si and P limitation, nitrogen (N) limitation restricts toxin production due to insufficient levels of free N to synthesize DA. In one study where *P. multiseries* was N-limited and failed to

produce DA, addition of nitrate subsequently stimulated DA production (Bates et al., 1991).

An inverse relationship has been demonstrated between DA production and growth rate of *P. multiseries* (Bates et al., 1996; Pan et al., 1996). This relationship has been attributed to the availability of high-energy intermediates necessary for DA synthesis, which varies over the growth cycle of *P. multiseries* (Pan et al., 1996, 1998). During exponential phase, cells are actively growing and less ATP is available for DA synthesis, whereas at stationary phase, carbon assimilation is reduced so available ATP may be used to support DA production (Pan et al., 1996).

Few laboratory studies have been completed in species other than *P. multiseries*. In *P. seriata*, the pattern of DA production was similar to that of *P. multiseries*, with minimal toxin production during exponential phase and increased production throughout stationary phase (Lundholm et al, 1994; Fehling et al., 2004). In contrast, toxin production in *P. australis* began during exponential phase and remained fairly constant during stationary phase (Garrison et al., 1992).

Axenic vs. nonaxenic cultures: Several bacterial isolates have been shown to enhance DA production by *P. multiseries* (Bates et al., 1995). While *P. multiseries* can produce DA in axenic cultures (Douglas and Bates, 1992; Douglas et al., 1993), reintroduction of bacteria to axenic cultures resulted in increased DA production by 2 to 115 fold (Bates et al., 1995). There is no evidence that bacteria in these cultures are capable of autonomous DA production (Gaudet, 2001; Bates et al., 2004), and the mechanism for enhanced DA production due to bacterial presence is uncertain. Bacterial numbers increase after the beginning of stationary phase of *P. multiseries*, corresponding with increased toxin production. However, axenic cultures also exhibit the characteristic increase in DA production during stationary phase. One suggested hypothesis for enhanced DA production in non-axenic cultures vs. axenic cultures is that the bacteria produce or regenerate precursor molecules necessary for DA production, rather than directly contributing to DA synthesis (Douglas and Bates, 1992; Bates, 1998).

Asexual vs. sexual reproduction: As a diatom, *P. multiseries* demonstrates a decrease in cell size during vegetative division. The mean cell length of a population of any diatom decreases over successive generations. Diatom frustules are composed of two valves that fit together like a petri dish, with one larger valve (epitheca) overlapping the smaller valve (hypotheca). Therefore, each mitotic division results in the formation of two differently sized daughter cells, one that is the same size as the parent and one that is slightly smaller (Round et al., 1990). In *P. multiseries*, an observed decrease in the capability to produce DA may coincide with the decrease in cell length (Bates et al., 1998), although not all isolates necessarily follow this trend (Dr. Stephen Bates, personal communication). Interestingly, cell deformities also tend to appear in *P. multiseries* cells after a certain period in culture (Villac, 1996; Bates et al., 1998).

Sexual reproduction restores the original, larger cell dimensions of *P. multiseries* and also appears to restore DA production in cultures that had experienced a reduction in DA production over time. Davidovich and Bates (1998) described the sexual reproductive cycles of *P. multiseries* as follows: pairing of parent cells of opposite mating types, gamete production, fusion of gametes to form zygotes, enlargement of auxospores, and formation of long, initial cells that usually produced higher levels of DA than the original parent cells.

Molecular Technology: While a considerable amount of research has been completed to investigate the biology of *P. multiseries*, the molecular characterization of this organism has been limited up to now. Further knowledge of the pathways that control the growth and physiology of *P. multiseries,* including toxin production, requires characterization of the genes that govern the regulation of these pathways. Therefore, this thesis project employed molecular techniques to identify and initiate characterization of actively expressed genes in *P. multiseries*.

Only a subset of all encoded genes is expressed in any given cell, and the levels and timing of gene expression determine the fate of individual cells. The central dogma

of molecular genetics describes gene expression as the process of DNA transcription into messenger RNA (mRNA), which is subsequently translated into functional protein. Since gene expression is initiated at the transcriptional level, gene discovery has often focused on studying gene expression by measuring mRNAs. Comparing the amount of specific mRNAs between two samples provides a mechanism to screen for genes that are turned on or off under defined physiological or environmental conditions.

Several techniques have been developed to analyze differentially expressed genes between two or more populations of nucleic acids. These comparative techniques include subtractive hybridization (Sagerstrom et al., 1997) and microarray technology (Brown, 1999; Schena et al., 1995, 1996; Shalon et al., 1996). In the subtractive hybridization approach, mRNA from the first cell type is converted to single-stranded complementary DNAs (ss cDNAs), which are then hybridized to an excess of all the mRNAs that are expressed in the second cell type. Genes that are expressed in both cell types will form cDNA/mRNA duplexes, while cDNA that is expressed in only the first cell type will be single-stranded and can then be separated from the duplexes by a number of methods. Subtractive hybridization is a relatively simple technique, which has been particularly useful in the identification of single significant mRNAs such as the isolation of T-cell receptor mRNAs by comparing gene expression profiles between T and B cells (Hedrick et al.,1984) and the identification of the myoD gene, a master regulator of muscle differentiation (Davis et al., 1987). Within marine ecology, suppressive, subtractive hybridization is currently being used to identify genes that are up-regulated in fish exposed to various environmental contaminants (Tsoi, 2004). Alternative protocols to standard subtractive hybridization to identify differentially expressed transcripts include representational difference analysis (RDA) and suppression PCR, which are PCR-based selection techniques. While all of these techniques provide a method for discovery of differentially expressed genes with high sensitivity, they do not allow the survey of a broad number of genes in a high-throughput mode.

Microarrays allow the monitoring of thousands of genes in parallel. The first step in construction of a cDNA microarray is to create a cDNA library from reverse transcription of mRNAs in cells or tissues of interest. Frequently, a subset of the cDNAs will be sequenced to begin to identify genes of interest and to verify the quality of the library. cDNA arrays are constructed by depositing thousands of amplified cDNAs onto glass microscope slides, with each cDNA represented as an independent spot on the array. The cDNA microarray is then hybridized to fluorescently labeled cDNA prepared by reverse transcription of mRNA isolated from two different populations of interest. Competitive hybridization of two samples labeled separately with Cy3 and Cy5, allows the ratio of mRNA abundance between the two samples to be compared for each individual cDNA on the microarray (Brown and Botstein, 1999). Microarray analysis applied within the field of phytoplankton ecology offers the potential to discover genes involved in ecologically relevant processes, such as toxin biosynthesis, population growth and bloom dynamics, photosynthesis, and nutrient cycling.

Microarray technology has proven to be a powerful tool for gene discovery programs in a wide range of organisms. In human cancer genetics, for example, microarray studies have led to the investigation of new approaches to diagnosis and drug therapy (Ochs and Goodwin, 2003). Microarrays have also been extremely useful in characterizing the transcriptional control mechanisms which govern physiological response in *S. cerevesiae* (Eisen et al., 1998; Spellman et al., 1998; Gasch et al., 2000). The success of mining large gene expression data sets and the potential for the information to be useful beyond the initial goals of this project suggested the application of microarray technology to the present study aimed at the identification of genes that are differentially expressed in *P. multiseries*. Advantages of DNA microarrays include 1) thousands of transcripts can be analyzed simultaneously 2) arrays allow simultaneous comparison of multiple samples, 3) a relatively small amount of starting material is required, 4) groups of genes with parallel expression patterns can be identified 5) the method is fast, efficient, and accurate, and 6) arrays can be useful for obtaining markers of specific physiological states.

cDNA microarrays have recently been applied to ecological studies. Genes associated with response to environmental variation and stress have been identified using microarrays in a number of studies. For example, one study identified cyanobacterial genes that were differentially expressed under conditions of high light acclimation, carbon dioxide fixation and photoprotection (Hihara et al., 2001). Other studies selected for cyanobacterial genes that responded rapidly to different wavelengths and intensities of irradiance (Huang et al. 2002; Gill et al., 2002). In the dinoflagellate *P. lunula*, microarray analysis has been successfully used to identify genes which are differentially expressed in relation circadian rhythm (Okamoto and Hastings, 2003). Microarray studies have also been applied directly to the analysis of environmental samples. For example, Taroncher-Oldenburg et al. (2003) addresses detrification in the Choptank River-Chesapeake Bay system using microarray methodology. Other studies demonstrated the effectiveness of microarray technology to monitor nitrogen cycling genes in environmental samples (Wu et al., 2001).

Summary:

While a considerable amount of research has been dedicated to understanding the biology of *P. multiseries*, many questions remain unanswered. A limitation to further knowledge of the biochemical pathways that control *Pseudo-nitzschia* physiology and growth, including domoic acid production, is posed by the lack of understanding of the molecular biology of this marine diatom. In general, diatoms have not received the same attention within the field of molecular biology that they have received in the fields of ecology and marine biology. However, the past few years have seen encouraging developments in the area of diatom genomics. Whole genome sequencing has recently been completed for the non-toxic, centric marine diatom *Thalassiosira pseudonana* (Armbrust et al., University of Washington, and US Department of Energy Joint Genome Institute). In addition, large-scale EST projects are currently being executed for *T. pseudonana* (Hildebrand et al., Scripps Institute of Oceanography, and US Dept of Energy Joint Genome Institute), and the non-toxic, pennate diatom *Phaeodactylum tricornutum* (Chris Bowler, Laboratory of Molecular Plant Biology, Stazione Zoologica).

The goals of this study were to establish a cDNA library and EST database for the toxic, pennate diatom *Pseudo-nitzschia multiseries* and to screen for differentially expressed genes using microarray technology. This approach was selected to identify and initiate characterization of genes associated with toxin production and the regulation of growth and physiology in this organism.

Chapter II

Pseudo-nitzschia multiseries cDNA Library and Expressed Sequence Tag Analysis

Abstract:

A complementary DNA (cDNA) library and expressed sequence tag (EST) database were constructed to identify and initiate characterization of actively expressed genes in the toxic marine diatom, *Pseudo-nitzschia multiseries*. A set of 3872 ESTs was generated by sequencing of 2552 randomly picked cDNA clones. 1320 cDNAs were sequenced in both the 3' and 5' directions, while 1232 cDNAs were sequenced in either the 3' or 5' direction. The ESTs were assembled into 1955 non-redundant contigs, of which 21% demonstrated significant similarity with known protein coding sequences.

The *P. multiseries* EST database included highly significant matches with sequences from all of the major taxonomic groups described within the universal phylogenetic tree. While some matches undoubtedly reflect the biases of the sequence databases, others likely reflect the evolutionary history of diatoms. Comparisons of the *P. multiseries* sequences against the *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* sequence databases proved useful in identifying diatom-specific transcripts. In addition, the discovery of numerous transcripts that did not match any known sequences in the public databases, nor any entry in the *T. pseudonana* and *P. tricornutum* databases offer novel sequences that will potentially help to elucidate unique aspects of *P. multiseries* biology, such as toxin production.

Key enzymes involved in C4 photosynthesis were revealed though sequence similarity, including a C4-specific pyruvate, orthophosphate dikinase, a phosphoenolpyruvate carboxykinase, a phosphoenolpyruvate carboxylase, and a pyruvate carboxylase. The existence of a C4 pathway in diatoms is currently under debate, so this discovery is particularly exciting, as it suggests the possibility of a C4 mechanism in *P. multiseries*. Many possible candidate genes that may play a role in DA biosynthesis were also revealed through sequence similarity to known protein coding sequences. Examples include enzymes involved in glutamate metabolism, such as 5-oxo-L-prolinase, acetylglutamate kinase, NAD-specific glutamate dehydrogenase, and N-acetylglutamate semialdehyde dehydrogenase.

Introduction:

Pseudo-nitzschia multiseries represents an ecologically important species within the marine phytoplankton. *P. multiseries* belongs to the division Bacillariophyta, unicellular brown algae commonly called diatoms, which contribute significantly to global carbon fixation (Hasle, 1995; Falkowski, 1998; Smetacek, 1999; Kooistra, 2003). Diatoms form the base of the food web in many marine environments and play a major role in nutrient cycling, especially in the biotransformation of silicon into silica during cell wall synthesis (Treguer et al., 1995). *P. multiseries* is distinctive among the marine diatoms, because it is one of the only known organisms to produce the neurotoxin, domoic acid (DA) (Bates et al, 1989, 1998; Todd, 1993). DA is a neuroexcitatory, water soluble amino acid, which has caused poisonings of humans, marine mammals, and birds though trophic transfer via shellfish consumption (Bates, 1989; Beltran, 1997; Scholin et al, 2000).

Despite its ecological importance, the molecular characterization of *P. multiseries* has been minimal. This is illustrated by the lack of protein-coding sequences available for *P. multiseries* in the public databases; a search of the updated NCBI database on August 8, 2004, yielded no entries for *P. multiseries*, and only four sequences for the related genuses *Nitzschia* and *Pseudo-nitzschia* combined. These sequences included malate dehydrogenase, which was characterized in the marine diatom *Nitzschia alba* (Yueh et al., 1989). The other three sequences likely encode 6-phosphogluconate dehydrogenase, cytochrome oxidase, and a delta-5 fatty acid desaturase, based on sequence similarity with known protein coding sequences (Ehara et al., 2000; Andersson and Roger, 2002). The lack of available information on the expressed genome of *P. multiseries* presents a limitation to further understanding the metabolic pathways that control cell physiology, including toxin production and growth. Therefore, a genomic program aimed at rapidly cataloguing actively expressed genes by sequencing complementary DNAs (cDNAs) was established as the most efficient initial approach to directly contribute to the expansion of this field using molecular genomics.

The sequencing and subsequent identification of cDNAs by similarity with known protein-coding sequences, called expressed sequence tag (EST) analysis, has become an important and well-established technique for gene discovery (Liang et al., 2000; Rudd, 2003). The basic strategy for EST analysis requires construction of a cDNA library from actively expressed mRNAs, followed by selection of cDNA clones at random to perform single, automated sequencing from one or both ends of the insert. The sequences are then categorized based on similarity to sequences deposited in public databases. This approach, which allows rapid assignment of function to a suite of actively expressed genes, is especially useful in organisms or tissues that previously have had little genetic inquiry or exploration. For example, the first application of high-throughput sequencing of cDNA clones allowed the isolation and subsequent characterization of numerous transcripts specific to the human brain (Adams et al., 1991).

The application of high-throughput sequencing of cDNA clones to investigate the biology of marine algae was introduced relatively recently with a study on the marine kelp, *Laminaria digitata* (Crepineau et al., 2000). At the onset of this thesis project, little information was available on diatom genomics, specifically. However, the past few years have resulted in an exciting opening of the field. A sequencing project on the marine diatom, *Phaeodactylum tricornutum*, has yielded a large EST dataset (Scala et al., 2002). While the original report described 1000 ESTs, a recent review of the updated sequence data available on-line revealed over 12,000 ESTs deposited from *P. tricornutum*. These ESTs were derived from the 5' end of the cDNAs and correspond to approximately 5100 non-redundant gene-oriented clusters (Chris Bowler, Laboratory of Molecular Plant Biology, Stazione Zoologica; http://avesthagen. sznbowler.com). The *P. tricornutum* project involves a multi-facility, interactive group that supports the data analysis and gene annotation of this large set of data and has now initiated the sequencing of the complete genome of *P. tricornutum.* Concurrently, an EST project on the marine diatom, *Thalassiosira pseudonana*, has yielded 17,000 ESTs, generated from 8500 cDNAs sequenced in both directions (Hildebrand et al., Scripps Institute of Oceanography, and US Dept of Energy Joint Genome Institute; http://avesthagen.sznbowler.com).

The *T. pseudonana* EST project supports the annotation of genes in the recently completed whole genome sequencing project on this diatom (Armbrust et al., University of Washington, and US Department of Energy Joint Genome Institute; http://genome.jgi-psf.org).

In the present study, a cDNA library and EST database were established for the toxic, pennate diatom, *Pseudo-nitzschia multiseries*. This project has currently generated 3872 ESTs, corresponding with 2552 cDNAs that were assembled into 1955 non-redundant contigs. The sequence information presented in this study will enable molecular tools to be further exploited in order to advance our understanding of the metabolic pathways that control the biology of *P. multiseries*. Comparative studies across the three diatom genomes should prove useful to the study of functional genomics and phylogeny among the diatoms.

Materials and Methods:

Culture conditions and RNA extraction: *Pseudo-nitzschia multiseries* clone CL-125 was graciously provided by Stephen S. Bates (Department of Fisheries and Oceans, Gulf Fisheries Center, Moncton, NB, Canada.) This clone was originally collected from Mill River, Prince Edward Island, Canada, on September 21, 2000, and isolated on September 23, 2000. Clonal cultures of CL-125 were grown in 0.2 μm filtered seawater enriched with f/2 nutrients (Guillard and Ryther, 1962). Batch cultures were maintained at 20 °C, 100 μE m^{-2} s^{-1}, 14:10 h LD cycle. Fifteen L of culture were grown in 19-L borosilicate carboys; the cultures were aerated using aquarium pumps with sterile tubing and were constantly mixed with magnetic stirrers. Cells were harvested during late exponential to mid-stationary growth phase, under predominantly toxin-producing conditions.

An extraction protocol which led to the consistent isolation of high-quality mRNA from *P. multiseries* was developed through the evaluation of a series of standard protocols for RNA isolation from other organisms. The final RNA extraction protocol given below yielded approximately 1 mg of total RNA and 10-12 μg of poly (A)+ RNA from 8 x 10^8 *P. multiseries* cells. *P. multiseries* cells were collected by centrifugation for 15 minutes at 1000*g*. Total RNA was extracted by homogenizing the cells (Polytron) in TRIzol (Invitrogen, Cat. No. 15596-018), which relies on lysis of the cells in the presence of both phenol and guanidium thiocyanate. Following homogenization, insoluble material was removed by low speed centrifugation of the samples, which increased both yield and quality of the resulting total RNA. Precipitating twice with salt and ethanol also contributed to high quality total RNA, as indicated by 260/280 O.D. ratios and gel electrophoresis. Poly (A)+ RNA was then isolated from total RNA using biotin-labeled oligo$(dT)_{20}$ probe bound to steptavidin magnetic particles. The method relies on poly (A) residues at the 3' ends of the mRNAs base-pairing with the oligo$(dT)_{20}$ probe. The bound polyadenylated RNA was magnetically isolated from the total RNA and purified (Roche, Cat. No. 1741985). Percent recovery of mRNA from total RNA was approximately 1-1.2%.

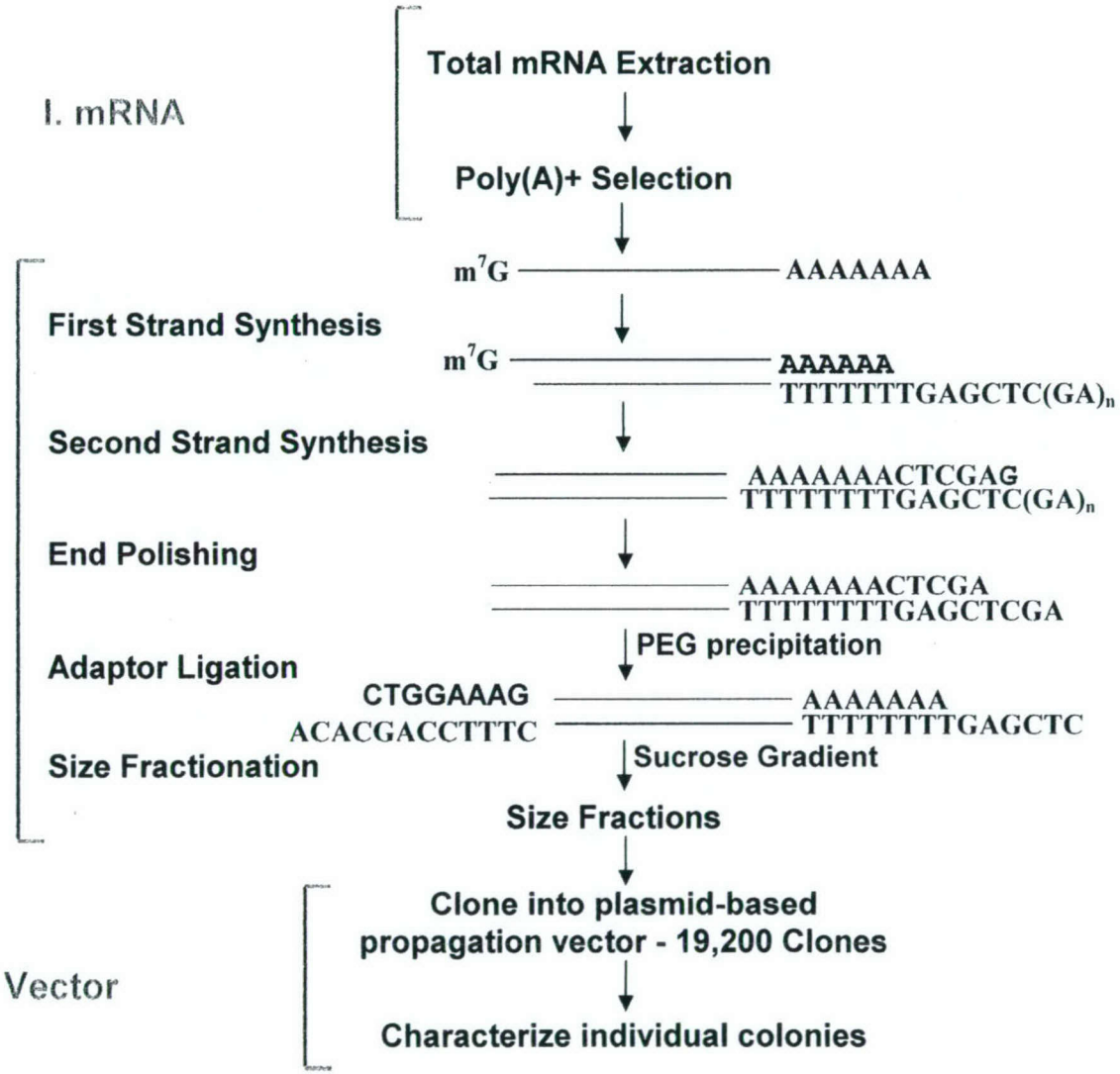
cDNA library (Figure 2-1): First-strand cDNA was prepared from poly (A) + RNA using Superscript II, NC-p7 (an RNA chaperone), and oligo pd(TZ) (an oligo (dT) primer with some of the internal thymidine residues replaced with 3-nitropyrrole to minimize mispriming to internal A-rich sequences). Double-stranded cDNA was generated using RNase H, *E. coli* DNA polymerase I, and *E. coli* ligase (to add a polymeric tract to the first-strand cDNA for initiation of second-strand synthesis) . The ends of the cDNA were polished with T4 DNA polymerase and GstXI adaptors were ligated to the cDNA ends. The cDNA was then fractionated on sucrose gradients. Individual size fractions were ligated into a pUC-based vector and transformed, by electroporation, into *E. coli* DH10B cells (Das et al., 2001). Following an initial library plating, individual colonies were picked and stored at -80°C in 15% glycerol for further analysis. Randomly chosen clones were then grown overnight in 1 mL of Terrific Broth. Plasmids were prepared for sequencing using an alkaline lysis method modified from Sambrook and Russell (2001). Alternatively, insert was amplified from randomly picked clones and then purified using Millipore multiscreens (see methods and materials in array section, Chapter 3).

Sequence Analysis: Sequence reactions were run on an automated DNA sequencer, ABI 3700 with dye terminators. The majority of the sequencing reactions were run in the laboratory of Jerry Pelletier, Biochemistry Department, McGill University. However, selected cDNAs from the expression studies presented in the next chapter were sequenced at ACGT, Inc. ESTs were edited to remove low quality data, poly (A) tails, and vector sequence. Automated trimming was performed using Seqman (DNAStar), followed by manual editing in order to proof-read and further remove low quality (ambiguous) data and poly (A) tails from the ends of the sequence. Vector sequence was removed using ContigExpress (VectorNTI). Vector removal was then verified by attempting to align vector sequence with the edited cDNA sequence in GenomeBench (VectorNTI.) The sequences were further edited by hand to remove any trace vector sequence revealed in this alignment process.

Figure 2-1:

Schematic diagram of cDNA library construction

Modified from Das et al., 2001



Multiple sequences from the same cDNA clone were assembled into consensus sequences using Seqman (DNAStar). Clone consensus sequences and singleton ESTs were further assembled to group the entire sequence dataset into unique classes of overlapping identical sequences, referred to as contigs (Cooke et al., 1997). A total of 1955 non-redundant consensus sequences were generated, using a criterion of 90% identity observed over sequences more than 50 nucleotides long. These parameters were based on a comparison of different criteria and software packages that revealed that Seqman (DNAStar) yielded the most consistent results using these limits, as demonstrated by the ability to group redundant sequences together consistently, without including non-related sequences. The final set of sequences will be deposited into the NCBI dbEST database.

Individual and consensus sequences were compared with known sequences contained within the public non-redundant protein databases using the Basic Local Alignment Search Tool provided by the NCBI server (Altschul et al., 1997; http://www.ncbi.nlm.nih.gov/BLAST/). Significant similarities were considered for E-values less than or equal to 7E-5. The E-value is a parameter that describes the number of hits that would be expected by chance; this value indicates the statistical significance of a given pairwise alignment. The lower the E-value, the more significant the hit. Specific *P. multiseries* sequences were also searched against the *Thalassiosira pseudonana* genome database at: http://genome.jgi-psf.org, and the *T. pseudonana* EST database and the *P. tricornutum*database at: http://avesthagen.sznbowler.com/. In these alignments, % identity and % similarity of the coding sequences compared were reported.

Results:

A cDNA library was constructed from *Pseudo-nitzschia multiseries* cells harvested during predominantly toxin-producing conditions, from late exponential to mid-stationary growth phase. A total of 19,200 cDNA clones were individually selected for growth and storage, after the initial library plating. The range of cDNA insert size was 500 to 4000bp, averaging 1000bp. A set of 2552 clones was randomly selected for sequencing. Of these, 1320 cDNAs were sequenced in both the 3' and 5' directions, while 1232 cDNAs were sequenced in either the 3' or 5' direction. Average sequence length for individual reads was 675bp, after vector removal and end-trimming (Table 2-1).

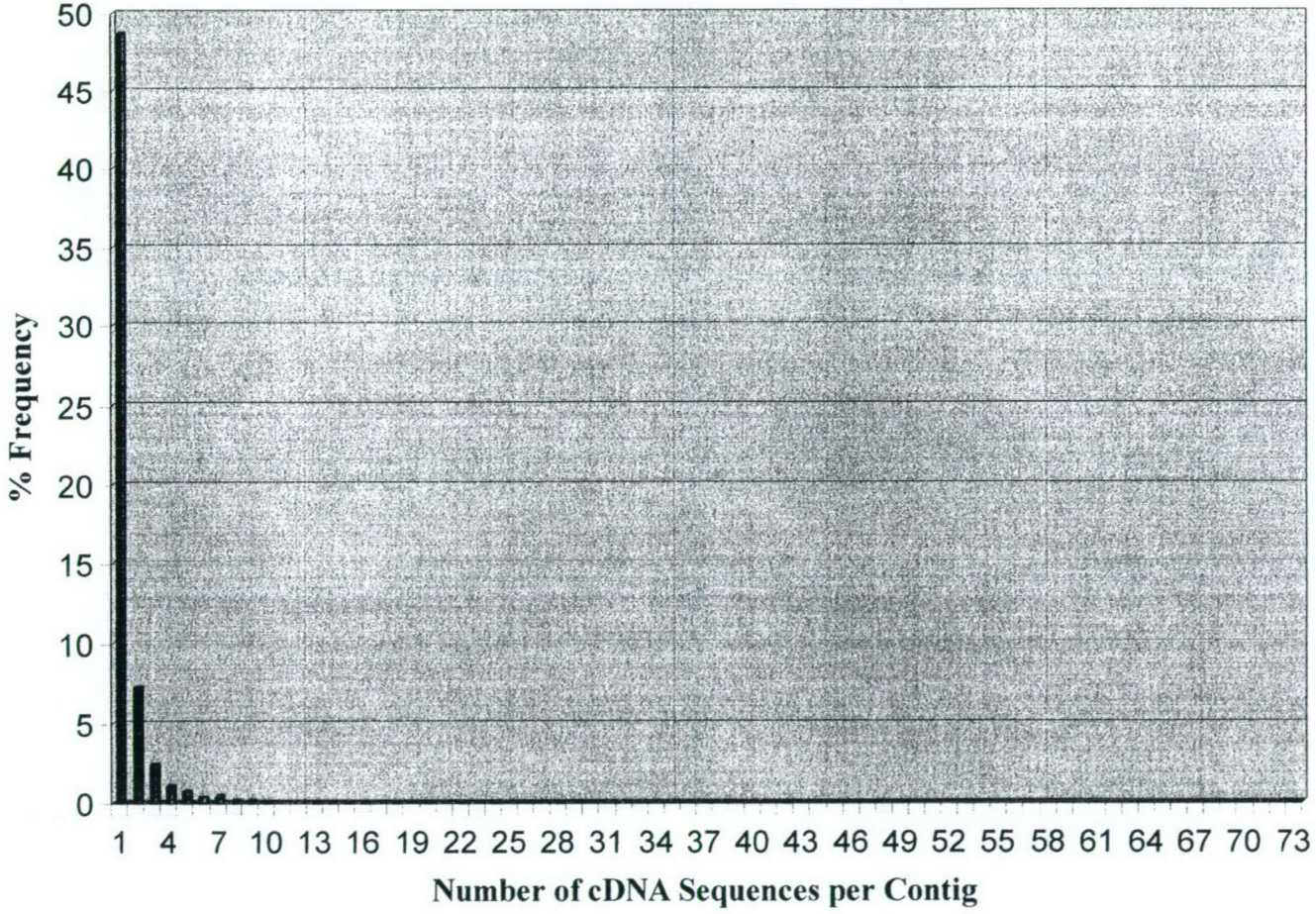
Assessment of library saturation, based on number of clones within each contig graphed against percent frequency, illustrated that total redundancy was relatively low (Figure 2-2). Of the 2552 cDNAs analyzed in this study, sequence assembly revealed 1955 represented non-redundant sequences or unique contigs, indicating a redundancy of 23.4%. The proportion of the *P. multiseries* cDNA library that appears in the sample of reads in this study may be approximately estimated by $C = 1 - n_1/n$, where n_1 is the number of genes that appear exactly once in the sampling and n is the total number of clones sequenced in this study (Susko and Roger, 2004). The expected number of reads required to discover a new gene may then be roughly estimated as $E = 1/(1-C)$. In this study, $n_1 = 1242$ and $n = 2552$. So, coverage in this analysis equals approximately 0.51, and the expected number of reads required to discover a new gene in this library would be 2.05. These estimates predict that further sequencing of this library would yield an additional 2000 unique transcripts. Therefore, including rare or low-copy transcripts, the *P. multiseries* cDNA library likely contains greater than 4000 expressed genes in total.

The *P. multiseries* deduced amino acid sequences were searched against the public non-redundant (nr) protein database, assigning a significant E-value of less than or equal to 7E-5 for 21.0% of the assembled consensus sequences against known proteins

Table 2-1. Pseudo-nitzschia multiseries cDNA Library /EST Summary

Total cDNA Clones Individually Selected for Growth and Storage, after Initial Library Plating	**19,200**
Range of Insert Size Based on both Plasmid Digest and Sequencing Results	**500-4000 bp (Est. avg. = 1 kb)**
Total Number of cDNAs Represented in EST Database	**2552**
Total Number of Unique Contigs or Non-redundant Sequences after Sequence Assembly	**1955**
Average Length of Sequence Read before End-trimming and Vector Removal	**798 bp**
Average Length of Sequence Read after End-trimming and Vector Removal	**675 bp**

Figure 2-2. Assessment of mRNA redundancy in the *P. multiseries* library by sequence assembly analysis. The number of individual cDNA clone sequences per contig is plotted against the percent frequency of the independent contigs.



(Table 2-2.) The *P. multiseries* cDNA sequences that demonstrated significant similarity to known protein coding sequences were categorized into functional groups, shown in Figure 2-3, while the putative identities of the individual non-redundant sequences that demonstrated significant similarity to known proteins are listed under functional group headings in Table 2-3.

In addition to known proteins, 3.7% of the *P. multiseries* EST database showed significant similarity to hypothetical sequences, and 2.7% showed significant similarity to unknown, environmental sequences. The unknown, environmental sequences were derived from a shotgun sequencing study in the nutrient replete Sargasso Sea (Venter et al., 2004). While this study targeted bacterial populations through size selection, the high sequence similarity with *P. multiseries* may indicate that their samples included eukaryotic algae, as well. Alternatively, the sequence similarity may reflect the evolutionary history of diatoms. In addition, some *P. multiseries* sequences with high similarity to known protein coding sequences also aligned with unknown sequences from the Sargasso Sea. For example, one environmental sequence aligned closely with a *P. multiseries* sequence that also showed high sequence similarity to the coding sequence for phosphoenolpyruvate carboxykinase, an enzyme involved in gluconeogenesis, anaplerotic reactions, and C4 photosynthesis (Lea et al., 2001). Characterization of *P. multiseries* sequences that are similar to unknown Sargasso Sea sequences may offer further understanding of the role that photosynthetic plankton play in the unique environment of the open ocean.

The *P. multiseries* EST database included highly significant matches for all of the major groups in the universal phylogenetic tree (Figure 2-4). While some hits against distant species may reflect the biases of the sequence databases, others likely reflect the evolutionary history of diatoms. Diatom lineages appear to have arisen through a secondary endosymbiosis between a heterotrophic flagellate that engulfed a single-celled red alga, which itself traces back to a primary endosymbiotic event in which a heterotrophic protist engulfed a cyanobacterium (Bhattacharya et al., 2003). Therefore,

Table 2-2. Overview of Blast results: Non-redundant *P. multiseries* sequences against NR protein database

	Number of Contigs	***% of Total***
High sequence similarity to known protein in NR database (E-value <7E-5)	**411**	**21.0%**
High sequence similarity to hypothetical protein in NR database (E-value <7E-5)	**73**	**3.7%**
High sequence similarity to unknown, environmental sequence in NR database (E-value <7E-5)	**53**	**2.7%**
Low sequence similarity to known protein in NR database (8E-5 to 1)	**504**	**25.8%**
Low sequence similarity to known protein in NR database (1 to 9.9)	**586**	**30.0%**
No Hits in NR database	**328**	**16.8%**
Total	**1955**	**100.00%**

Figure 2-3: Functional classification of derived coding sequences from *Pseudo-nitzschia multiseries*

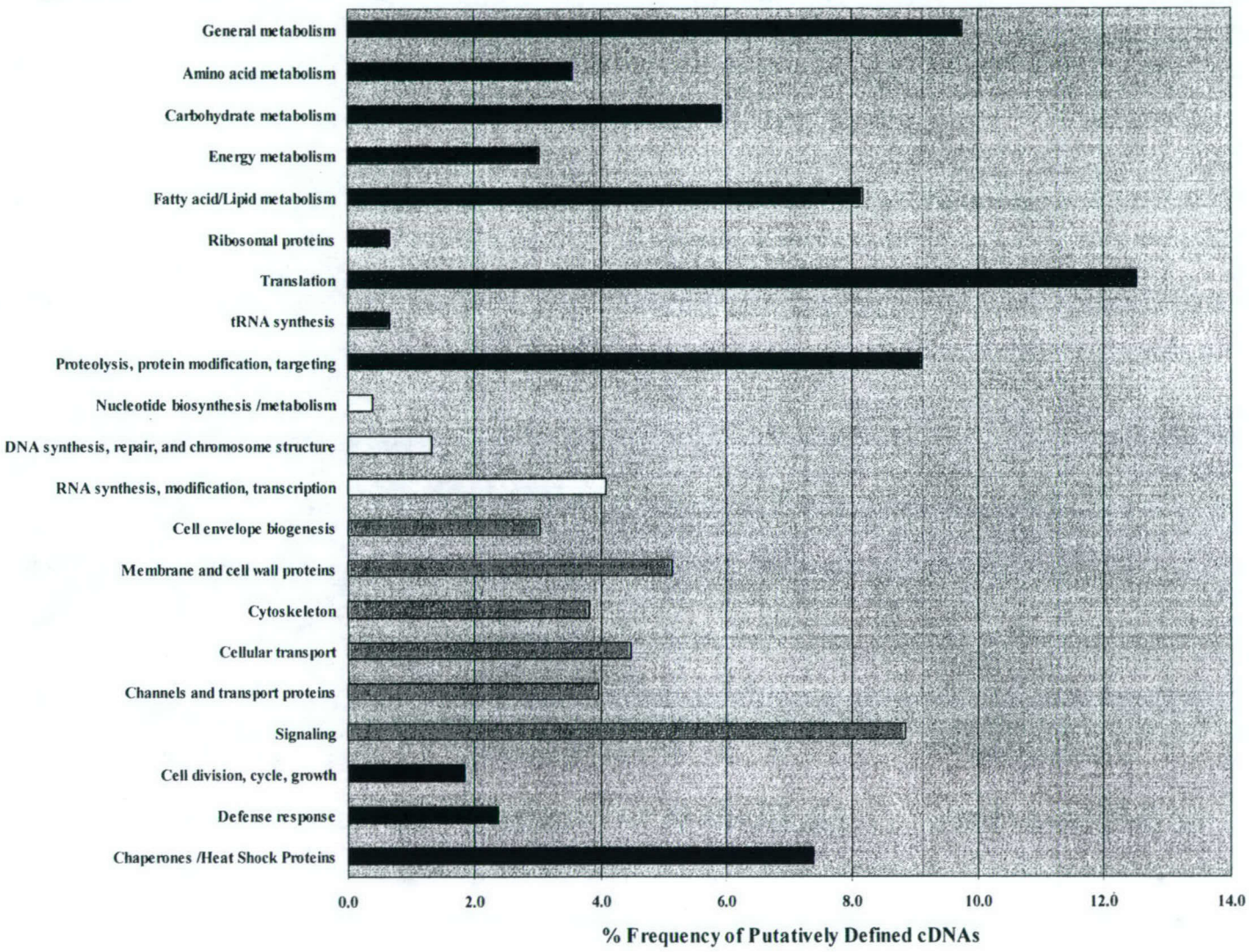


Table 2-3. Non-redundant consensus sequences from the *Pseudo-nitzschia multiseries* cDNA library that demonstrated significant similarity to known proteins in NCBI's protein database.

PSN I.D.	cDNAs per Contig	Sequence Length (bp)	NCBI Accession No.	Putative Identification	*Species* or Domain Name	E-value
1. Cell metabolism						
a. General (74)						
PSN0011	22	2263	ZP_00064353.1	3-carboxymuconate cyclase	COG2706	8.00E-48
54C11	1	720	BAD13433.1	5-oxo-L-prolinase	*Bos taurus*	2.00E-52
50H11	1	374	ZP_00033423.1	Acetamidase/formamidase, predicted	*Burkholderia fungorum*	1.00E-26
PSN767	9	1044	AAO89237.1	Adenosylhomocysteinase	*Arabidopsis thaliana*	e-105
PSN0015	8	1742	NP_200170.2	Aldo/keto reductase family protein	*Arabidopsis thaliana*	3.00E-47
166D3	1	707	AAL47846.1	Aldose reductase	*Candida boidinii*	3.00E-42
166E12	1	810	ZP_00089447.1	Alpha/beta hydrolase superfamily	*Azotobacter vinelandii*	5.00E-06
164C5	1	737	ZP_00034033.1	Alpha/beta hydrolase superfamily	*Burkholderia fungorum*	2.00E-21
55B7	1	623	AAC26842.1	Alpha-N-acetylglucosaminidase	*Mus musculus*	1.00E-23
174F11	1	646	BAA83575.1	Arginine N-methyl transferase 1, putative	*Oryza sativa*	2.00E-34
50B4	1	252	NP_896399.1	ATP-sulfurylase	*Synechococcus sp.*	9.00E-13
174H9	1	780	AAF20208.1	Cingulin	*Xenopus laevis*	1.00E-06
PSN0090	2	1676	T02955	Cytochrome P450 monooxygenase, probable	*Zea mays*	2.00E-15
186E12	1	825	AAF67724.2	Cytochrome P450, defense	*Diabrotica virgifera*	5.00E-16
169C11	1	790	O22101	Ferrochelatase II	*Oryza sativa*	3.00E-52
7B8	1	717	NP_773877.1	GTP cyclohydrolase II, putative	*Bradyrhizobium japonicum*	5.00E-53
47G4	1	719	ZP_00205201.1	Heme iron utilization protein, putative	*Pseudomonas aeruginosa*	3.00E-18
50A9	1	337	NP_565876.2	Heme-binding family protein, SOUL	*Arabidopsis thaliana*	4.00E-09
179A10	1	782	AAM64677.1	HesB protein	*Arabidopsis thaliana*	3.00E-16
PSN466	3	1546	AAM64493.1	Hydroxymethyltransferase	*Arabidopsis thaliana*	e-123
PSN596	2	1220	AAM64677.1	Iron-sulfur cluster assembly complex protein	*Arabidopsis thaliana*	7.00E-13
53F5	1	1393	NP_191712.1	Ketopantoate hydroxymethyltransferase family	*Arabidopsis thaliana*	5.00E-56

50G5	1	730	NP_595123.1	Mannosyltransferase, probable	*Schizosaccharomyces pombe*	2.00E-07
PSN289	2	842	NP_869842.1	Mercuric reductase	*Pirellula sp. 1*	5.00E-16
PSN1714	2	1145	NP_177495.2	MutT/nudix family protein (Hydrolase)	*Arabidopsis thaliana*	2.00E-10
55A6	1	515	AAN34969.1	ov-thioredoxin 1	*Onchocerca volvulus*	3.00E-09
24D8	1	745	NP_627664.1	Oxidoreductase, putative	*Streptomyces coelicolor*	7.00E-10
183G7	1	755	NP_828006.1	Pyridoxine biosynthesis protein, putative	*Streptomyces avermitilis*	7.00E-28
183A4	1	784	ZP_00175543.2	RTX toxins and related Ca2+-binding proteins	*Crocosphaera watsonii*	1.00E-08
A1D1	1	785	ZP_00163759.1	SAM-dependent methyltransferases	*Synechococcus elongatus*	2.00E-11
53D7	1	209	AAL35384.1	Serine hydroxymethyltransferase	*Chlamydomonas reinhardtii*	7.00E-05
45C11	1	600	T08094	Sulfate adenylyltransferase, putative	*Chlamydomonas reinhardtii*	6.00E-65
57A6	1	569	Q98TX1	Thioredoxin	*Ophiophagus hannah*	1.00E-14

b. Amino acid metabolism (27)

30C10	1	716	NP_487562.1	2-isopropylmalate synthase	*Nostoc sp.*	8.00E-41
45B4	1	607	ZP_00212771.1	Acetylglutamate kinase	*Burkholderia cepacia*	2.00E-10
7F6	1	746	NP_201379.2	Branched-chain amino acid aminotransferase 5	*Arabidopsis thaliana*	1.00E-26
51B11	1	781	P08955	CAD: Glutamine carbamoyl-phosphate synthase	*Mesocricetus auratus*	2.00E-42
55C8	1	985	NP_445914.1	Cysteine desulfurase	*Rattus norvegicus*	e-103
164E8	1	801	NP_508168.1	Cysteine dioxygenase	*Caenorhabditis elegans*	3.00E-15
164D2	1	790	AAM18799.1	Diaminopimelate decarboxylase	*Dictyostelium discoideum*	9.00E-09
53A1	1	578	NP_770397.1	Glutathione S-reductase	*Bradyrhizobium japonicum*	7.00E-29
PSN1239	2	774	P46436	Glutathione S-transferase	*Ascaris suum*	2.00E-10
165B8	1	829	NP_947833.1	N-acetylglutamate semialdehyde dehydrogenase	*Rhodopseudomonas palustris*	2.00E-44
PSN1428	3	1369	ZP_00131198.1	NAD-specific glutamate dehydrogenase	*Desulfovibrio desulfuricans*	4.00E-39
PSN918	2	808	T50771	Peptidylprolyl isomerase	*Solanum tuberosum*	8.00E-12
PSN969	2	840	AAP82284.1	Phenylalanine hydroxylase	*Danio rerio*	2.00E-44
16A3	1	743	BAD07294.1	Prolyl 4-hydroxylase	*Nicotiana tabacum*	2.00E-08

164E6	1	707	AAN31489.1	S-adenosyl methionine synthetase	*Phytophthora infestans*	2.00E-59
54B10	1	702	ZP_00163553.1	S-adenosylmethionine methyltransferase, predicted	*Synechococcus elongatus*	4.00E-25
174C4	1	760	NP_181225.1	S-adenosylmethionine synthetase, putative	*Arabidopsis thaliana*	8.00E-65
47C3	1	557	ZP_00146777.2	Serine-pyruvate aminotransferase	*Psychrobacter sp.*	2.00E-13
165E6	1	794	NP_296185.1	Serine-pyruvate aminotransferase	*Deinococcus radiodurans*	4.00E-46
55D3	1	709	ZP_00141212.1	Spermidine synthase	*Pseudomonas aeruginosa*	4.00E-10
52G7	1	767	BAB86769.1	Tauropine dehydrogenase	*Arabella iricolor*	2.00E-20
45E4	1	774	NP_594579.1	Threonine synthase	*Schizosaccharomyces pombe*	4.00E-61

c. Carbohydrate metabolism (45)

30E2	1	611	ZP_00059817.1	3-phosphoglycerate kinase	*Clostridium thermocellum*	2.00E-35
16H11	1	603	AAL76320.1	6-phosphogluconate dehydrogenase	*Phytophthora infestans*	2.00E-57
136A7	1	569	AAD30364.1	celB endoglucanase	*Caldicellulosiruptor sp.*	6.00E-06
183D11	1	793	BAA33802.1	Cytosolic phosphoglycerate kinase 1	*Populus nigra*	4.00E-27
163G6	1	935	NP_182103.1	Eukaryotic phosphomannomutase family protein	*Arabidopsis thaliana*	4.00E-85
136A3	1	694	AAP79192.1	Fructose-1,6 bisphosphatase	*Bigelowiella natans*	8.00E-31
75G12	1	1278	ZP_00022038.1	Fructose-2,6-bisphosphatase	*Ralstonia metallidurans*	2.00E-09
PSN1130	2	687	NP_281780.1	Fructose-bisphosphate aldolase	*Campylobacter jejuni*	1.00E-32
PSN1138	5	830	AAF34327.1	Glyceraldehyde-3-phosphate dehydrogenase	*Odontella sinensis*	9.00E-74
45H1	1	511	AAF34325.1	Glyceraldehyde-3-phosphate dehydrogenase, cytosolic	*Phaeodactylum tricornutum*	2.00E-42
185E8	1	784	NP_819776.1	KpsF/GutQ family protein	*Coxiella burnetii*	2.00E-25
PSN0016	14	2158	NP_282084.1	Phosphoenolpyruvate carboxykinase (ATP)	*Campylobacter jejuni*	e-164
174A10	1	814	NP_908415.1	Phosphoenolpyruvate carboxylase 2	*Oryza sativa*	7.00E-20
164H10	1	770	BAB71853.1	Phosphoenolpyruvate carboxylase kinase	*Flaveria trinervia*	2.00E-21
183E7	1	773	AAF45021.1	Phosphoglycerate kinase precursor	*Phaeodactylum tricornutum*	2.00E-98
175A9	1	883	NP_661303.1	Phosphoglycerate mutase	*Chlorobium tepidum*	4.00E-53
186D10	1	782	NP_952663.1	Phosphoglycerate mutase	*Geobacter sulfurreducens*	3.00E-51
PSN0100	2	919	NP_869505.1	PPi-phosphofructokinase	*Pirellula sp. 1*	4.00E-67
PSN1264	2	833	NP_571625.1	Pyruvate carboxylase	*Danio rerio*	2.00E-31

PSN0103	2	1726	BAA21654.1	Pyruvate orthophosphate dikinase **C4	*Eleocharis vivipara*	e-143
24E1	1	668	NP_193288.1	Pyruvate phosphate dikinase family protein	*Arabidopsis thaliana*	6.00E-48
24B6	1	439	P46285	Sedoheptulose-1,7-bisphosphatase	*Triticum aestivum*	7.00E-23
183B1	1	453	NP_621883.1	Thiamine pyrophosphate dehydrogenases	*Thermoanaerobacter sp.*	7.00E-10
51C7	1	1289	BAC16227.1	Xylitol dehydrogenase	*Gluconobacter oxydans*	1.00E-48

d. Energy metabolism (23)

164H12	1	833	AAR84642.1	AAA ATPase	*Candida albicans*	1.00E-41
50B11	1	609	NP_181830.1	Amine oxidase family protein	*Arabidopsis thaliana*	3.00E-07
PSN332	2	765	AAH60012.1	ATPase family, AAA domain	*Xenopus laevis*	3.00E-36
50C5	1	642	O80433	Citrate synthase, mitochondrial precursor	*Daucus carota*	9.00E-41
179E11	1	663	P00567	Creatine kinase	*Oryctolagus cuniculus*	3.00E-17
54E2	1	1092	AAK69398.1	Cytochrome b5 reductase PF36	*Cucurbita maxima*	2.00E-29
56B11	1	703	AAN31477.1	Electron transfer flavoprotein beta subunit	*Phytophthora infestans*	5.00E-53
16G8	1	785	NP_442688.1	Ferredoxin	*Synechocystis sp.*	3.00E-09
54D9	1	555	NP_894633.1	Flavodoxin	*Prochlorococcus marinus*	2.00E-32
135H6	1	975	AAB80924.1	Fucoxanthin-chlorophyll a/c light-harvesting protein	*Skeletonema costatum*	2.00E-50
51B6	1	781	AAP80710.1	Light-harvest protein	*Griffithsia japonica*	4.00E-14
56B3	1	767	CAC87422.1	Light-harvesting protein	*Galdieria sulphuraria*	4.00E-15
50A3	1	635	AAR20479.1	Mitochondrial cytochrome c peroxidase	*Cryptococcus neoformans*	3.00E-24
57G2	1	545	NP_991386.1	NADH dehydrogenase (ubiquinone) 42 kDa subunit	*Bos taurus*	4.00E-22
PSN1717	2	782	AAN77914.1	NADH dehydrogenase subunit 1	*Homo sapiens*	1.00E-78
50F8	1	261	NP_113453.1	Photosystem II stability/assembly factor HCF136	*Guillardia theta*	4.00E-24
179E1	1	645	YP_005802.1	Quinone oxidoreductase	*Thermus thermophilus HB27*	2.00E-20
PSN0847	2	693	1SOX	Sulfite Oxidase, A Chain A	*Gallus gallus*	1.00E-29
51F12	1	730	T48590	Ubiquinol-cytochrome-c reductase	*Arabidopsis thaliana*	7.00E-33
56H3	1	696	BAB41213.1	Zeta-crystallin (NADPH:quinone reductase)	*Hyla japonica*	8.00E-15

e. Fatty acid/Lipid metabolism (62)

17B10	1	878	NP_037266.2	3-hydroxy-3-methylglutaryl-Coenzyme A reductase	*Rattus norvegicus*	1.00E-66
53F9	1	424	P77851	3-Hydroxybutyryl-CoA Dehydrogenase	*Thermoanaerobacterium sp.*	2.00E-08
174G9	1	793	NP_294792.1	3-hydroxybutyryl-CoA dehydrogenase	*Deinococcus radiodurans*	2.00E-14
16A9	1	541	AAH45119.1	Acetyl-Coenzyme A acyltransferase 2	*Xenopus laevis*	8.00E-34
PSN0081	4	2727	NP_638218.1	Acyl-CoA dehydrogenase	*Xanthomonas campestris*	e-117
PSN0829	2	791	NP_491859.1	Acyl-Coenzyme A dehydrogenase short branched chain	*Caenorhabditis elegans*	3.00E-19
45H7	1	783	ZP_00005474.1	Acyl-coenzyme A synthetases/AMP-(fatty) acid ligases	*Rhodobacter sphaeroides*	2.00E-66
PSN0073	2	1451	NP_250308.1	AMP-binding enzyme, probable	*Pseudomonas aeruginosa*	9.00E-24
174G6	1	758	BAB32665.1	Branched-chain alpha-keto acid dehydrogenase E1	*Gallus gallus*	1.00E-14
172G11	1	515	NP_990387.1	CFR-associated protein p70	*Gallus gallus*	6.00E-32
PSN0057	5	1668	AAL92563.1	Delta 6 fatty acid desaturase D6	*Phaeodactylum tricornutum*	0
51E8	1	622	AAD40245.1	Delta-9-stearoyl-acyl carrier protein desaturase, plastid	*Brassica juncea*	2.00E-06
53F6	1	804	AAO16600.1	Digalactosyldiacylglycerol synthase	*Xerophyta humilis*	1.00E-16
56E11	1	666	NP_295210.1	Enoyl-CoA hydratase/3,2-trans-enoyl-CoA isomerase	*Deinococcus radiodurans*	3.00E-29
25H12	1	647	XP_227370.2	Farnesyl-pyrophosphate synthetase	*Rattus norvegicus*	2.00E-52
170C5	1	889	NP_446059.1	Fatty acid Coenzyme A ligase, long chain 5	*Rattus norvegicus*	1.00E-09
169B8	1	770	AAK11525.1	Geranylgeranyl pyrophosphate synthase	*Penicillium paxilli*	1.00E-61
183B4	1	736	AAO72312.1	L-3-hydroxyacyl-CoA dehydrogenase subunit precursor	*Euglena gracilis*	7.00E-37
47F10	1	555	NP_014405.1	Lecithin cholesterol acyl transferase (LCAT)	*Saccharomyces cerevisiae*	7.00E-14
165H11	1	858	AAA61791.1	Lipoxygenase	*Porphyra purpurea*	3.00E-14
PSN0037	6	2101	ZP_00084847.1	Long-chain acyl-CoA synthetases (AMP-forming)	*Pseudomonas fluorescens*	1.00E-95
PSN0054	3	2216	ZP_00187590.1	Long-chain acyl-CoA synthetases (AMP-forming)	*Rubrobacter xylanophilus*	1.00E-66

PSN0079	2	278	ZP_00044073.1	Long-chain acyl-CoA synthetases (AMP-forming)	*Magnetococcus sp.*	5.00E-08
PSN0014	13	2438	NP_662047.1	Long-chain-fatty-acid-CoA ligase	*Chlorobium tepidum*	4.00E-48
45E6	1	1182	AAH38229.1	Neuropathy target esterase	*Homo sapiens*	4.00E-38
16B6	1	615	NP_666363.2	Neuropathy target esterase, related	*Mus musculus*	1.00E-12
136C3	1	672	NP_009121.1	Phospholipase (Sec23-interacting protein p125)	*Homo sapiens*	5.00E-09
53B6	1	530	XP_358347.1	Phospholipase A2, group IVB (cytosolic)	*Mus musculus*	4.00E-05
166A7	1	815	NP_948772.1	Salicylate hydroxylase, possible	*Rhodopseudomonas palustris*	9.00E-06
PSN0617	2	1000	AAN77732.1	Stearoyl-CoA desaturase	*Oreochromis mossambicus*	2.00E-41
PSN1320	2	766	AAR87713.1	Stearoyl-CoA desaturase	*Sus scrofa*	5.00E-18

2. Protein Metabolism

a. Ribosomal proteins (5)

163A1	1	1011	NP_182283.1	60S ribosomal protein L7A (RPL7aA)	*Arabidopsis thaliana*	4.00E-56
30E6	1	483	P42037	Ribosomal P2, acidic 60S ribosomal protein	*Alternaria alternata*	6.00E-12
5G12	1	914	AAK92160.1	Ribosomal protein L22	*Spodoptera frugiperda*	4.00E-28
165F6	1	541	AAD47076.1	Ribosomal protein L8	*Anopheles gambiae*	2.00E-39
167H9	1	765	AAN05595.1	Ribosomal protein S8	*Argopecten irradians*	9.00E-64

b. Translation (95)

51G1	1	780	ZP_00083483.1	Asp-tRNAAsn/Glu-tRNAGln amidotransferase A subunit	*Pseudomonas fluorescens*	7.00E-16
53H11	1	762	AAH02841.1	Dimethyladenosine transferase (rRNA methylation)	*Homo sapiens*	8.00E-57
PSN0001	73	1628	AAK27413.1	Elongation factor 1alpha long form	*Monosiga brevicollis*	e-115
52F6	1	404	NP_702375.1	Elongation factor 2 (EF-2)	*Plasmodium falciparum*	2.00E-48
57A4	1	735	XP_323573.1	Elongation factor 2 kinase EFK-1B isoform, related	*Neurospora crassa*	7.00E-10
169D12	1	573	BAA33895.1	Elongation factor 3	*Yarrowia lipolytica*	9.00E-42
PSN1720	2	546	AAC35391.1	Elongation-like factor	*Candida albicans*	1.00E-20
PSN1327	2	1185	NP_700577.1	Eukaryotic translation initiation factor 2, beta, putative	*Plasmodium falciparum*	9.00E-30

50D3	1	437	P24922	Eukaryotic translation initiation factor 5A-2	*Nicotiana plumbaginifolia*	1.00E-29
16C1	1	986	NP_196751.2	Eukaryotic translation initiation factor SUI1 family protein	*Arabidopsis thaliana*	8.00E-09
PSN0845	3	1133	NP_892518.1	Light repressed protein A, homolog	*Prochlorococcus marinus*	1.00E-21
47E3	1	346	CAA41267.1	Mitochondrial elongation factor G	*Saccharomyces cerevisiae*	1.00E-21
47C2	1	735	CAA41267.1	Mitochondrial elongation factor G	*Saccharomyces cerevisiae*	1.00E-19
PSN1484	2	794	NP_492457.1	Translation Elongation Factor (94.8 kD) (eft-2)	*Caenorhabditis elegans*	1.00E-68
53H8	1	827	NP_594081.1	Translation initiation factor 2 alpha subunit, Eukaryotic	*Schizosaccharomyces pombe*	1.00E-62
55A12	1	718	CAC43441.1	Translation initiation factor 4A, Eukaryotic	*Toxoplasma gondii*	4.00E-68
165A5	1	466	AAM94013.1	Translation initiation factor 6	*Griffithsia japonica*	2.00E-14
164F2	1	732	Q8YEB3	Translation initiation factor IF-2	*Brucella melitensis*	7.00E-28

c. tRNA synthesis (5)

183C2	1	760	EAK89327.1	Glycyl-tRNA synthetase	*Cryptosporidium parvum*	3.00E-34
165B6	1	843	NP_497837.1	Leucyl tRNA Synthetase	*Caenorhabditis elegans*	1.00E-37
51C3	1	681	EAK89712.1	Methionyl-tRNA synthetase	*Cryptosporidium parvum*	4.00E-22
179C6	1	637	NP_594867.1	Seryl-tRna synthetase, cytoplasmic	*Schizosaccharomyces pombe*	3.00E-23
177D3	1	724	ZP_00015568.1	Tyrosyl-tRNA synthetase	*Rhodospirillum rubrum*	6.00E-13

d. Proteolysis, protein modification, targeting (69)

177G2	1	537	Q9FXT9	26S proteasome AAA-ATPase subunit RPT1	*Oryza sativa*	4.00E-65
PSN0031	15	1980	BAB86297.1	Alkaline serine protease IV	*Alteromonas sp. O-7*	2.00E-46
47B11	1	644	CAC86003.1	Aspartic proteinase	*Theobroma cacao*	7.00E-07
PSN0496	3	1556	AAA74445.1	Cathepsin B protease, putative	*Urechis caupo*	2.00E-24
PSN0739	4	1569	AAC60301.1	Cathepsin D	*Oncorhynchus mykiss*	2.00E-51
53A3	1	724	AAG28507.1	Cathepsin Z	*Mus musculus*	1.00E-10
136E5	1	775	AAC39839.1	Cathepsin Z precursor; CTSZ	*Homo sapiens*	2.00E-17
186D1	1	925	NP_033926.1	Cysteine protease (calpain)	*Mus musculus*	7.00E-61
16B9	1	551	AAG45422.1	E3 ubiquitin ligase SMURF2	*Homo sapiens*	4.00E-23
57C3	1	795	NP_188548.2	Peptidase M16 /Insulinase family protein	*Arabidopsis thaliana*	5.00E-25
54B12	1	1051	ZP_00202329.1	Periplasmic protease	*Synechococcus elongatus*	4.00E-41

54B1	1	485	AAC35858.1	Polyubiquitin	*Capsicum chinense*	5.00E-12
25E2	1	650	AAP80690.1	Polyubiquitin	*Griffithsia japonica*	2.00E-23
PSN0032	8	1409	NP_701482.1	Polyubiquitin, PfpUB	*Plasmodium falciparum*	0
PSN0595	4	1148	CAF32070.1	Prohibitin, putative	*Aspergillus fumigatus*	8.00E-62
PSN0203	2	1246	Q9CR00	Proteasome 26S non-ATPase regulatory subunit 9	*Mus musculus*	3.00E-07
57D2	1	744	Q9CR00	Proteasome 26S non-ATPase regulatory subunit 9	*Mus musculus*	3.00E-07
24C4	1	419	S17521	Proteasome endopeptidase complex	*Homo sapiens*	8.00E-09
54C1	1	691	NP_915931.1	Proteasome subunit alpha type 3	*Oryza sativa*	1.00E-45
50E4	1	799	O64464	Proteasome subunit beta type 1	*Oryza sativa*	2.00E-28
50E3	1	1045	NP_036095.1	Proteasome subunit, alpha type 1	*Mus musculus*	1.00E-59
165B5	1	807	BAA10932.1	Proteosome LMPX	*Petromyzon marinus*	3.00E-56
PSN0113	2	1264	AAQ76845.1	Serine carboxypeptidase CBP1	*Trypanosoma cruzi*	1.00E-69
52B8	1	700	AAO85509.1	SGT1, ubiquitin ligase	*Nicotiana benthamiana*	5.00E-21
164G7	1	521	AAB69653.1	Trypsinogen 1	*Boltenia villosa*	2.00E-06
7E6	1	850	AAF00920.1	Ubiquitin	*Oxytricha trifallax*	9.00E-94
PSN0603	4	1380	AAB88617.1	Ubiquitin conjugating enzyme	*Zea mays*	6.00E-12
50H1	1	770	NP_973557.1	Ubiquitin fusion degradation UFD1 family protein	*Arabidopsis thaliana*	2.00E-36
25A2	1	611	AAL32102.1	Ubiquitin ligase E3 alpha-II	*Mus musculus*	2.00E-15
137C10	1	922	AAM76730.1	Ubiquitin ligase NEDD4h	*Homo sapiens*	6.00E-57
PSN0116	2	881	NP_594902.1	Ubiquitin ligase, putative	*Schizosaccharomyces pombe*	6.00E-24
47G6	1	685	P35135	Ubiquitin-conjugating enzyme E2-17 kDa	*Lycopersicon esculentum*	2.00E-16
54E4	1	798	CAA66655.1	Ubiquitin-protein ligase, E3 (E6-AP)	*Homo sapiens*	2.00E-32
53G12	1	826	NP_193887.1	Vacuolar protein sorting-associated protein 28	*Arabidopsis thaliana*	1.00E-36

3. Nucleic Acid Metabolism

a. Nucleotide biosynthesis /metabolism (3)

167E10	1	818	ZP_00127094.1	Glutamine phosphoribosylpyrophosphate amidotransferase	*Pseudomonas syringae*	1.00E-22
54G6	1	979	YP_005603.1	Inosine-5'-monophosphate dehydrogenase	*Thermus thermophilus HB27*	9.00E-08
7H11	1	912	NP_980755.1	Uridine kinase	*Bacillus cereus*	8.00E-43

b. DNA synthesis, repair, and chromosome structure (10)

56H5	1	764	NP_296978.1	Helicase, putative	*Chlamydia muridarum*	3.00E-20
184B12	1	983	NP_296978.1	Helicase, putative	*Chlamydia muridarum*	2.00E-19
30A4	1	592	T06392	Histone H1	*Lycopersicon esculentum*	7.00E-06
177F1	1	705	CAB82768.1	Histone H3	*Fucus serratus*	8.00E-32
51E5	1	781	AAC60120.1	Inner Centromere protein, XL-INCENP	*Xenopus laevis*	6.00E-05
50D1	1	779	NP_064550.2	Postreplication repair protein hRAD18p; RAD18	*Homo sapiens*	1.00E-06
PSN1712	3	763	0611193A	Replicatory Protein P	*Bacteriophage lambda*	7.00E-78
51C1	1	663	NP_201476.1	SNF2 domain-containing protein	*Arabidopsis thaliana*	8.00E-36

c. RNA synthesis, modification, transcription (31)

PSN1139	2	1120	NP_593520.1	ATP dependent RNA helicase, putative	*Schizosaccharomyces pombe*	1.00E-20
164D6	1	750	T48796	ATP-dependent RNA helicase DED1, probable	*Neurospora crassa*	3.00E-76
169C6	1	769	BAC79194.1	Chloroplast nucleoid DNA-binding protein, related	*Oryza sativa*	5.00E-24
52A6	1	460	NP_938179.1	DEAD (Asp-Glu-Ala-Asp) box polypeptide 49	*Danio rerio*	2.00E-29
177C8	1	746	AAL87139.2	DEAD box RNA helicase Vasa	*Cyprinus carpio*	5.00E-30
57E9	1	673	NP_587856.1	DEAD/DEAH box RNA helicase	*Schizosaccharomyces pombe*	1.00E-58
30H12	1	754	CAA67363.1	High mobility group (HMG)	*Lampetra fluviatilis*	3.00E-20
165D6	1	802	NP_194111.1	High mobility group (HMG1/2) family protein	*Arabidopsis thaliana*	3.00E-05

183C8	1	466	NP_011617.1	Nucleolar protein involved in rRNA processing	*Saccharomyces cerevisiae*	1.00E-20
171C1	1	942	NP_191917.2	Nucleotidyltransferase family protein	*Arabidopsis thaliana*	1.00E-05
136B2	1	575	XP_323582.1	Regulator of nonsense transcripts, putative	*Neurospora crassa*	1.00E-35
174B12	1	782	T48246	Ribonuclease II-like protein	*Arabidopsis thaliana*	4.00E-05
179E7	1	694	NP_057280.1	RNA binding motif protein 19	*Homo sapiens*	8.00E-15
PSN1150	2	586	NP_060531.1	RNA binding protein, putative	*Homo sapiens*	7.00E-18
57H3	1	685	AAK18841.1	RNA polymerase I subunit, putative	*Oryza sativa*	2.00E-14
PSN1407	4	1208	NP_568143.1	SNF7 family protein	*Arabidopsis thaliana*	1.00E-21
55H3	1	908	NP_565336.1	SNF7 family protein	*Arabidopsis thaliana*	1.00E-20
177A8	1	777	NP_173115.1	TAZ zinc finger family protein / zinc finger (ZZ type) family	*Arabidopsis thaliana*	1.00E-30
PSN0435	3	2107	AAK20748.1	ToxR-activated gene A protein	*Vibrio cholerae*	2.00E-39
PSN1348	4	883	NP_177329.2	Transducin family protein / WD-40 repeat family	*Arabidopsis thaliana*	8.00E-10
54F6	1	614	AAL92018.1	UPF1 (Up-frameshift suppressor 1)	*Arabidopsis thaliana*	2.00E-56

4. Cell Structure and Function

a. Cell envelope biogenesis (23)

PSN0201	4	1222	AAF13033.2	Beta(1-3)endoglucanase	*Aspergillus fumigatus*	1.00E-31
PSN0067	9	1447	NP_248049.1	Capsular polysaccharide biosynthesis protein I	*Methanocaldococcus sp.*	4.00E-72
25D8	1	762	AAK30369.1	GalNAc-4-sulfotransferase 2	*Homo sapiens*	1.00E-05
PSN1240	2	801	NP_198236.1	NAD-dependent epimerase/dehydratase family	*Arabidopsis thaliana*	2.00E-52
PSN0414	3	1495	ZP_00085359.1	Nucleoside-diphosphate-sugar epimerases	*Pseudomonas fluorescens*	2.00E-07
PSN0095	3	1437	NP_924014.1	Nucleotide sugar epimerase	*Gloeobacter violaceus*	2.00E-60
7D9	1	1148	NP_997192.1	UDP-N-acteylglucosamine pyrophosphorylase 1-like 1	*Homo sapiens*	2.00E-57

b. Membrane and cell wall proteins (39)

7A12	1	814	A60610	Circumsporozoite protein precursor	*Plasmodium brasilianum*	1.00E-24
PSN0756	2	1136	Q03650	Cysteine-rich, acidic integral membrane protein precursor	*Trypanosoma brucei*	9.00E-19

PSN0959	3	1461	NP_563881.1	Endomembrane protein 70, putative	*Arabidopsis thaliana*	6.00E-94
25H8	1	448	NP_001414.1	Epithelial membrane protein 1	*Homo sapiens*	3.00E-57
186B3	1	774	AAM64352.1	ER lumen retaining receptor (HDEL receptor), putative	*Arabidopsis thaliana*	7.00E-51
PSN0184	7	1071	CAA06854.1	F-spondin	*Branchiostoma floridae*	4.00E-16
25D4	1	641	CAA06854.1	F-spondin	*Branchiostoma floridae*	7.00E-15
45G2	1	739	NP_113896.1	Glycoprotein 110; cell membrane; adhesion regulating	*Rattus norvegicus*	2.00E-08
50G10	1	755	NP_869081.1	Membrane associated protein, putative	*Pirellula sp.*	5.00E-05
16C7	1	706	NP_493807.1	Membrane protein family member (2B70), putative	*Caenorhabditis elegans*	6.00E-05
52A4	1	810	ZP_00162262.2	Membrane protein, predicted	*Anabaena variabilis*	5.00E-05
53E2	1	782	ZP_00164190.2	Membrane protein, predicted	*Synechococcus elongatus*	7.00E-05
PSN0755	5	1667	AAO73245.1	Membrane protein, putative	*Oryza sativa*	4.00E-19
76A6	1	903	NP_687759.1	Membrane protein, putative	*Streptococcus agalactiae*	4.00E-10
16E6	1	777	NP_197476.1	Peroxisomal membrane 22 kDa family protein	*Arabidopsis thaliana*	8.00E-07
45D8	1	573	NP_974505.1	Peroxisomal membrane protein-related	*Arabidopsis thaliana*	2.00E-13
45D3	1	953	NP_975632.1	Prolipoprotein	*Mycoplasma mycoides subsp. mycoides SC str. PG1*	3.00E-09
PSN0068	4	1067	AAO51196.1	Synaptobrevin-like protein	*Dictyostelium discoideum*	1.00E-08
PSN0301	2	1400	EAK88414.1	TB2/DP1/HVA22 family integral membrane protein	*Cryptosporidium parvum*	1.00E-07
185C1	1	560	EAK88414.1	TB2/DP1/HVA22 family integral membrane protein	*Cryptosporidium parvum*	2.00E-14
135A11	1	934	NP_442911.1	Transforming growth factor induced protein	*Synechocystis sp.*	3.00E-05
186H3	1	760	NP_975970.1	Variable surface prolipoprotein, putative	Mycoplasma mycoides	9.00E-25

c. Cytoskeleton (29)

PSN0019	7	1584	P26182	Actin	*Achlya bisexualis*	0
PSN1161	3	1102	P26182	Actin	*Achlya bisexualis*	e-117
186F1	1	834	S49007	Actin	*Pythium irregulare*	e-103
136F3	1	628	P26182	Actin	*Achlya bisexualis*	4.00E-67
47B12	1	274	BAB62395.1	Actin	*Nannochloris coccoides*	5.00E-09

PSN1090	3	1133	AAG01044.1	Actin	*Pythium splendens*	e-140
179B5	1	292	AAO14682.1	Actin	*Pyrocystis lunula*	2.00E-11
177F11	1	694	EAA15967.1	Actin 3	*Plasmodium yoelii*	1.00E-18
161G10	1	245	S68090	Actin 8	*Arabidopsis thaliana*	7.00E-05
47G10	1	767	AAA51619.1	Actin 8	*Dictyostelium discoideum*	9.00E-11
167H5	1	1679	AAC47528.1	Actin-binding protein fragmin P	*Physarum polycephalum*	2.00E-39
177D10	1	691	NP_005727.1	Actin-related protein 1 homolog A	*Homo sapiens*	4.00E-38
25H11	1	513	CAD79598.1	Beta-tubulin	*Suberites domuncula*	7.00E-79
45E3	1	727	P34036	Dynein heavy chain, cytosolic (DYHC)	*Dictyostelium discoideum*	9.00E-28
25F1	1	603	NP_171954.1	Myosin	*Arabidopsis thaliana*	3.00E-18
52H11	1	476	AAN75607.1	Myosin 2	*Cryptococcus neoformans var. neoformans*	6.00E-15
179A6	1	761	NP_476934.2	Myosin CG9155-PD	*Drosophila melanogaster*	2.00E-25
PSN0929	2	960	A85318	Myosin heavy chain-like protein, imported	*Arabidopsis thaliana*	2.00E-18

d. Cellular transport (34)

6F1	1	331	NP_973722.1	Archain 1-like	*Danio rerio*	7.00E-10
PSN1093	2	798	P25870	Clathrin heavy chain	*Dictyostelium discoideum*	2.00E-52
PSN0318	4	973	AAG50828.1	Clathrin heavy chain, putative	*Arabidopsis thaliana*	6.00E-49
45E11	1	1364	CAE45585.1	Coatomer alpha subunit-like protein	*Lotus corniculatus*	8.00E-35
30A9	1	669	P53619	Coatomer delta subunit (Delta-coat protein)	*Bos primigenius*	4.00E-43
164A2	1	778	BAB02664.1	Coatomer protein complex, beta prime	*Arabidopsis thaliana*	6.00E-84
54G3	1	789	NP_176393.1	Coatomer protein complex, subunit alpha	*Arabidopsis thaliana*	2.00E-65
186B2	1	783	NP_004757.1	Coatomer protein complex, subunit beta 2	*Homo sapiens*	8.00E-52
179B11	1	683	NP_057212.1	Coatomer protein complex, subunit gamma 1	*Homo sapiens*	4.00E-14
24B9	1	615	AAM65018.1	Coatomer-like protein, epsilon subunit	*Arabidopsis thaliana*	2.00E-15
45G4	1	747	BAA75463.1	COP-coated vesicle membrane protein P24 homolog	*Polysphondylium pallidum*	9.00E-07
54E8	1	739	AAB72116.2	Importin	*Arabidopsis thaliana*	3.00E-61
54B11	1	760	AAP31033.1	Importin alpha	*Toxoplasma gondii*	3.00E-25
57C6	1	618	CAB40789.1	Importin alpha-3	*Drosophila melanogaster*	5.00E-15
169G11	1	766	1QGK	Importin Beta	*Homo sapiens*	2.00E-25
78B2	1	867	AAH54537.1	Kif4 protein, kinesin	*Mus musculus*	6.00E-08
52C11	1	751	BAC56912.1	Kinesin motor protein	*Dictyostelium discoideum*	2.00E-26

165F9	1	812	AAC05386.1	Nucleoporin	*Drosophila melanogaster*	1.00E-07
57F3	1	782	NP_013412.1	Protein involved in vesicular transport	*Saccharomyces cerevisiae*	1.00E-06
169C8	1	776	NP_567217.1	Protein transport, related	*Arabidopsis thaliana*	1.00E-06
7F1	1	551	1FVF	Sec1	*Loligo pealei*	5.00E-15
45E10	1	450	AAQ97847.1	Sec13-like protein	*Danio rerio*	5.00E-07
167H2	1	793	AAH45117.1	Sec61a1-prov protein	*Xenopus laevis*	e-102
PSN1423	2	1127	CAC03439.2	SNAP-beta	*Homo sapiens*	4.00E-11
PSN1238	3	1254	AAB72112.1	Vacuolar sorting receptor homolog	*Arabidopsis thaliana*	1.00E-30
169B1	1	759	T00044	Vacuolar sorting receptor protein homolog PV72	*Cucurbita sp.*	2.00E-27
53E10	1	839	AAF24062.1	v-SNARE AtVTI1b	*Arabidopsis thaliana*	1.00E-06

e. Channels and transport proteins (30)

PSN102	2	1640	AAL85708.1	ABC transporter ABCC.5	*Dictyostelium discoideum*	7.00E-98
165G6	1	789	NP_195847.1	ABC transporter family protein	*Arabidopsis thaliana*	1.00E-27
PSN1727	2	774	T04442	ABC-type transport protein, RNase L inhibitor-like	*Arabidopsis thaliana*	1.00E-82
PSN0682	2	1394	AAK26384.1	ADP/ATP carrier	*Toxoplasma gondii*	3.00E-56
53F12	1	678	NP_977232.1	Amino acid permease family protein	*Bacillus cereus*	3.00E-26
179D9	1	805	AAP92715.1	Calcium-transporting ATPase 1	*Ceratopteris richardii*	2.00E-23
7E1	1	619	NP_057032.2	CGI-19 protein, solute carrier family	*Homo sapiens*	3.00E-21
25F7	1	665	NP_704591.1	E1-E2 ATPase/hydrolase, putative	*Plasmodium falciparum 3D7*	3.00E-07
PSN0475	2	796	NP_505467.2	Golgi GDP-fucose translocator	*Caenorhabditis elegans*	4.00E-20
177F9	1	681	NP_850136.1	Inorganic phosphate transporter, putative	*Arabidopsis thaliana*	6.00E-08
30H5	1	546	BAB97617.1	Mg/Co/Ni transporter MgtE (contains CBS domain)	*Corynebacterium sp.*	1.00E-07
52C3	1	805	AAH43834.1	Mitochondrial Ca2+-dependent solute carrier	*Xenopus laevis*	3.00E-11
169F12	1	713	NP_565682.1	Mitochondrial import inner membrane translocase, TIM10	*Arabidopsis thaliana*	3.00E-08
PSN1722	2	1254	XP_209204.2	Mitochondrial inner membrane, transport, hypothetical	*Homo sapiens*	7.00E-11
164D9	1	189	AAC49653.1	SIT1 (silicon transporter)	*Cylindrotheca fusiformis*	3.00E-09
16F7	1	678	NP_865825.1	Sodium/sulfate symporter	*Pirellula sp. 1*	1.00E-07

PSN0072	3	1818	NP_009162.1	Solute carrier family 6, amino acid transporter BO+	*Homo sapiens*	8.00E-34
52G3	1	290	ZP_00047927.2	Sulfate permease and related transporters	*Magnetospirillum sp.*	1.00E-07
51F8	1	701	NP_187151.1	Transport protein particle (TRAPP) component Bet3 family protein	*Arabidopsis thaliana*	3.00E-11
56H6	1	735	NP_682730.1	Transporter, similar	*Thermosynechococcus sp.*	2.00E-08
78F8	1	756	Q43362	Vacuolar ATP synthase 16 kDa proteolipid subunit	*Pleurochrysis carterae*	3.00E-53
179H7	1	795	AAB49621.1	Vacuolar ATPase 100 kDa subunit	*Dictyostelium discoideum*	2.00E-36
171C7	1	873	BAC66648.1	Vacuolar membrane ATPase subunit a precursor	*Candida glabrata*	3.00E-83

f. Signaling (67)

169C5	1	774	AAG17931.1	Aardvark (beta-catenin)	*Dictyostelium discoideum*	4.00E-08
PSN0055	4	1758	NP_446247.1	Ankyrin rich membrane-spanning, KinaseD substrate	*Rattus norvegicus*	1.00E-14
PSN0078	3	1414	BAC79897.1	Ankyrin-kinase, putative	*Oryza sativa*	5.00E-16
186D7	1	789	NP_594320.1	azr1 protein (transports azoles across membrane)	*Schizosaccharomyces pombe*	3.00E-11
PSN0528	3	1727	AAS55542.1	Ca2+ and calmodulin-dependent protein kinase	*Pisum sativum*	2.00E-07
171A9	1	846	T37321	Ca2+/calmodulin-dependent protein kinase I	*Caenorhabditis elegans*	1.00E-46
170A2	1	800	BAC19848.1	Calcium/calmodulin-dependent protein kinase I alpha	*Xenopus laevis*	9.00E-06
183D6	1	785	CAA73906.1	Calmodulin; calcium-binding protein	*Ciona intestinalis*	2.00E-24
A1F6	1	401	AAA64341.1	cAMP-dependent protein kinase	*Gonyaulax polyedra*	7.00E-24
25H3	1	434	AAD12740.1	cAMP-regulated guanine nucleotide exchange factor I	*Homo sapiens*	1.00E-09
57C12	1	385	AAK62411.1	Casein kinase II alpha subunit	*Arabidopsis thaliana*	2.00E-14
169C1	1	789	EAK88911.1	DHHC family palmitoyl transferase	*Cryptosporidium parvum*	6.00E-21
179C12	1	806	NP_031970.1	Epidermal growth factor receptor pathway substrate 15	*Mus musculus*	3.00E-15
53A6	1	701	AAF73141.1	GTP-ase regulator RPGR	*Canis familiaris*	3.00E-11
PSN0255	4	1096	Q01890	GTP-binding protein ypt1	*Phytophthora infestans*	1.00E-78

47H7	1	751	YP_005324.1	Guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase	*Thermus thermophilus*	1.00E-22
7H8	1	764	AAN85440.1	Inositol 5-phosphatase 4	*Dictyostelium discoideum*	6.00E-23
PSN0966	2	1323	NP_502057.1	Kinase D-interacting substance of 220 kDa like (4L811)	*Caenorhabditis elegans*	2.00E-05
54D12	1	800	BAB41205.1	Kinase-like protein	*Oryza sativa*	1.00E-14
177H8	1	1267	NP_849538.1	Leucine-rich repeat protein kinase family protein	*Arabidopsis thaliana*	1.00E-22
53G4	1	777	NP_192625.3	Leucine-rich repeat protein kinase family protein	*Arabidopsis thaliana*	2.00E-11
PSN1390	2	839	NP_200956.1	Leucine-rich repeat transmembrane protein kinase, put.	*Arabidopsis thaliana*	5.00E-09
PSN0471	2	1131	NP_199445.1	Leucine-rich repeat transmembrane protein kinase, put.	*Arabidopsis thaliana*	2.00E-06
PSN0060	6	1975	EAK90389.1	Mucin, large thr stretch, signal peptide sequence	*Cryptosporidium parvum*	5.00E-06
16A10	1	595	CAB64999.2	Nuclear protein kinase 2, putative	*Euplotes octocarinatus*	1.00E-08
45D7	1	781	XP_166254.4	Odd Oz/ten-m homolog 4	*Homo sapiens*	5.00E-12
55F11	1	498	NP_178075.1	Protein kinase family protein	*Arabidopsis thaliana*	1.00E-08
186E10	1	269	A57676	Protein kinase Xa21	*Oryza sativa*	3.00E-07
56E2	1	765	NP_191885.1	Protein kinase, putative (MRK1)	*Arabidopsis thaliana*	7.00E-09
PSN1309	2	205	AAM88379.1	Protein phosphatase type 1 catalytic subunit gamma	*Canis familiaris*	6.00E-20
PSN0537	2	986	BAB83049.1	Protein tyrosine phosphatase	*Pyrococcus abyssi*	8.00E-06
50F4	1	782	NP_700999.1	Protein tyrosine phosphatase, putative	*Plasmodium falciparum 3D7*	5.00E-15
A1B4	1	506	NP_49123.	RAB family member (23.6 kD) (rab-2)	*Caenorhabditis elegans*	2.00E-34
A1C1	1	578	BAB97381.1	rab GDP-dissociation inhibitor	*Branchiostoma belcheri*	9.00E-48
167C1	1	811	CAB46230.1	rab GDP-dissociation inhibitor	*Branchiostoma floridae*	7.00E-46
55C10	1	808	BAC41349.1	Receptor protein Notch1	*Cynops pyrrhogaster*	4.00E-13
PSN1549	2	562	NP_914843.1	Receptor-like protein	*Oryza sativa*	9.00E-16
166A8	1	1204	ZP_00170241.2	Secreted and surface protein with fasciclin-like repeat	*Ralstonia eutropha*	3.00E-15
56B6	1	750	ZP_00170241.2	Secreted and surface protein with fasciclin-like repeat	*Ralstonia eutropha*	1.00E-09

55E8	1	1250	NP_567375.1	Serine/threonine protein phosphatase PP1 isozyme 6	*Arabidopsis thaliana*	e-105
137G1	1	990	BAC79157.1	Serine/threonine-protein kinase ctr1, putative	*Oryza sativa*	3.00E-46
164B11	1	792	ZP_00152402.2	Signal transduction histidine kinase	*Dechloromonas aromatica*	2.00E-08
7D5	1	798	NP_033457.1	TPR-containing, SH2-binding phosphoprotein	*Mus musculus*	2.00E-33
183H7	1	673	AAR37838.1	Twin-arginine translocation domain protein	*Uncultured bacteria*	2.00E-12
7F4	1	507	NP_486915.1	Two-component hybrid sensor and regulator	*Nostoc sp. PCC 7120*	9.00E-09
174H10	1	795	NP_195052.1	WD-40 repeat family protein	*Arabidopsis thaliana*	2.00E-15

5. Cell division, cycle, growth (14)

171C6	1	891	AAR98743.1	ASPM (mitotic spindle function)	*Aotus sp. PDE-2004*	5.00E-05
PSN0096	2	845	P41210	Caltractin (Centrin)	*Atriplex nummularia*	4.00E-06
16D3	1	753	P41210	Caltractin (Centrin)	*Atriplex nummularia*	4.00E-06
136C7	1	804	NP_189146.1	Cell division control protein, related	*Arabidopsis thaliana*	6.00E-05
PSN1255	2	784	NP_958857.1	Cell division cycle 27	*Danio rerio*	4.00E-68
30D4	1	673	AAN75620.1	CID1 (nucleotidyltransferase; S-M checkpoint)	*Cryptococcus neoformans*	2.00E-05
169H4	1	748	NP_567243.1	Cullin family protein	*Arabidopsis thaliana*	5.00E-46
169A11	1	767	AAL78999.1	Cyclin fold protein 1 variant b	*Homo sapiens*	3.00E-14
25G9	1	835	NP_006692.1	DIM1 (mitosis) homolog, Thioredoxin-like 4	*Homo sapiens*	2.00E-62
25H5	1	314	NP_056979.1	Geminin, DNA replication inhibitor	*Homo sapiens*	3.00E-41
165F7	1	811	BAB85218.1	Matrix metalloproteinase	*Volvox carteri*	1.00E-08
51H8	1	736	S71192	Mitosis-specific cyclin 2	*Arabidopsis thaliana*	4.00E-05

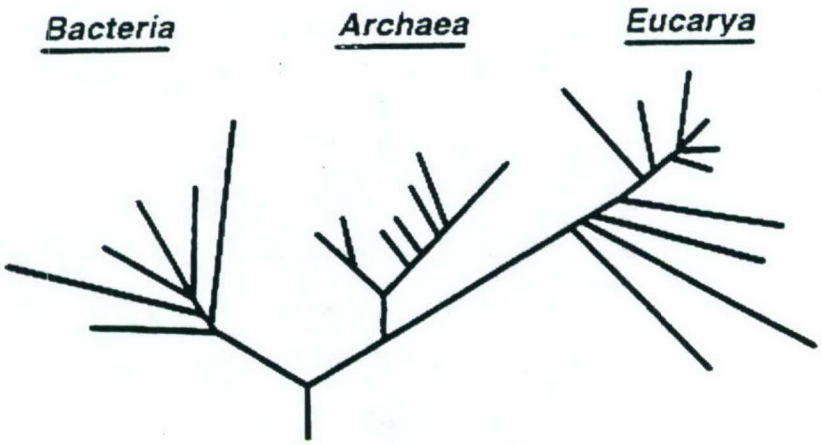
6. Defense response (18)

PSN0520	2	767	AAC78591.1	Disease resistance Cf-5	*Lycopersicon esculentum*	7.00E-19
167G5	1	827	T10504	Disease resistance protein Cf-2.1	*Lycopersicon sp.*	4.00E-30
164C3	1	698	T30553	Disease resistance protein Hcr2-5D	*Lycopersicon esculentum*	2.00E-14
25G11	1	1536	AAC78593.1	Disease resistance protein, Hcr2-0B	*Lycopersicon esculentum*	8.00E-28
169F9	1	794	AAC78595.1	Hcr2-5B	*Lycopersicon esculentum*	4.00E-12
53H3	1	1265	AAN17454.1	Hypersensitive-induced reaction protein 4	*Hordeum vulgare subsp.*	1.00E-51
PSN0089	2	1142	AAF68391.1	Hypersensitive-induced response protein, HIR3	*Zea mays*	8.00E-30

7F12	1	772	NP_921887.1	Hypersensitive-induced response protein, putative	*Oryza sativa*	2.00E-41
54E7	1	535	CAD45029.1	NBS-LRR disease resistance protein, homologue	*Hordeum vulgare*	8.00E-13
PSN0661	7	1371	NP_973504.1	Salicylic acid induction deficient 1 (SID1)	*Arabidopsis thaliana*	9.00E-19

7. Chaperones /Heat Shock Proteins (56)

PSN0753	5	1947	AAA80655.1	BiP	*Phaeodactylum tricornutum*	e-104
52D3	1	1268	AAM12857.1	Chaperonin containing TCP-1 delta subunit	*Physarum polycephalum*	e-109
57G3	1	791	AAM12858.1	Chaperonin containing TCP-1 epsilon subunit	*Physarum polycephalum*	2.00E-62
56E6	1	797	AAL56960.1	Chaperonin subunit alpha	*Malawimonas jakobiformis*	2.00E-46
57E8	1	684	NP_006727.2	DnaJ (Hsp40) homolog, subfamily B, member 2	*Homo sapiens*	6.00E-15
30E5	1	1076	EAK89499.1	DnaJ protein	*Cryptosporidium parvum*	3.00E-12
165H7	1	803	AAM93962.1	DnaJ protein	*Griffithsia japonica*	3.00E-08
PSN0020	10	1227	NP_113882.1	Heat shock factor 2	*Rattus norvegicus*	8.00E-10
PSN0458	6	1084	AAF01280.1	Heat shock protein 101	*Triticum aestivum*	9.00E-56
186E2	1	790	NP_186949.1	Heat shock transcription factor 2	*Arabidopsis thaliana*	3.00E-16
16B8	1	565	AAO66547.1	Heat shock transcription factor, putative	*Oryza sativa*	2.00E-07
16F2	1	721	AAF74563.1	Heat stress transcription factor A3	*Lycopersicon peruvianum*	1.00E-12
7D4	1	731	NP_193510.1	HSF1	*Arabidopsis thaliana*	1.00E-08
PSN0025	7	1001	HMMHSP20	Hsp 20 /alpha crystaliln	HSP20 SEED	9.70E-23
PSN0610	3	1083	Q00043	Hsp 70	*Ajellomyces capsulatus*	1.00E-78
PSN1076	7	1525	BAA97566.1	Hsp 70	*Blastocystis hominis*	e-162
PSN1011	2	808	AAR21576.1	Hsp 70	*Phytophthora nicotianae*	5.00E-26
54C10	1	812	BAC67671.1	Hsp 90	*Cyanidioschyzon merolae*	4.00E-32
PSN0554	2	1826	AAC28921.1	Hsp 90-1	*Achlya ambisexualis*	9.00E-65
52H3	1	789	AAR12194.1	Hsp 90-2	*Nicotiana benthamiana*	3.00E-40
167D7	1	780	NP_071909.1	SIL1, ER chaperone, BiP-associated protein	*Homo sapiens*	3.00E-05
50A2	1	585	NP_996761.1	T-complex protein 1 delta subunit	*Gallus gallus*	6.00E-57



Universal Phylogenetic Tree modified from Woese, 2000. This tree is derived from the phylogenetic comparison of rRNA sequences, and indicates 3 major domains. Multi-gene phylogenetic comparisons have further modified the current understanding of relationships within the Eukarya (see figure 2-5).

	A	B
Eukarya (80.8%)		
Fungi	**30**	**7.3%**
Metazoa	**110**	**26.8%**
Amoebozoa	**14**	**3.4%**
Plantae	**119**	**29.0%**
Rhodophyta	**7**	**1.7%**
Heterokonta	**20**	**4.9%**
Alveolata	**21**	**5.1%**
Euglenoids	**3**	**0.7%**
Excavata	**1**	**0.2%**
Bacteria (18.7%)		
Proteobacteria	**34**	**8.3%**
Cyanobacteria	**15**	**3.6%**
Other bacteria	**28**	**6.8%**
Archaea (0.5%)		
Archaea	**2**	**0.5%**

Figure 2-4. The *P. multiseries* EST database included highly significant matches for all of the major groups in the universal phylogenetic tree. The number (A) and percentage (B) of *P. multiseries* non-redundant sequences that showed significant sequence similarity to protein coding sequences from a species within the given group are presented.

diatoms would be expected to have genes deriving from both photosynthetic and heterotrophic lineages. The *P. multiseries* deduced amino acid sequences aligned most closely to eukaryotic proteins for 80.8% of the significant matches, with an almost even split between heterotrophic and phototrophic organisms. Similarity to bacterial sequences accounted for 18.7% of the *P. multiseries* sequences, of which 3.6% were of cyanobacterial origin and 8.3% were of proteobacterial origin. Proteobacteria are believed to be most closely related to the ancestral bacterial cell that led to mitochondria in eukaryotes (Gray et al., 1999). Only 2 archaeal sequences demonstrated significant similarity with the *P. multiseries* sequences. One of these corresponded to a capsular polysaccharide biosynthesis protein, with an E-value of 4E-72, 46% identity, and 64% similarity. Searching the *T. pseudonana* genome produced several similar sequences with up to 74% identity, 89% similarity to the *P. multiseries* sequence, and 42% identity, 62% similarity to the archaeal sequence (E-value, 2E-71). The *P. tricornutum* EST database did not produce a similar sequence. The absence of this sequence from the *P. tricornutum* EST database could be the consequence of low expression levels of the orthologous transcript. In contrast, the *P. multiseries* library contained at least 9 copies of this transcript, and the *T. pseudonana* genome appeared to contain four closely related sequences. Capsular polysaccharide biosynthesis proteins are involved in cell membrane biogenesis and signaling (Roberts, 1996). Therefore, this transcript may represent the discovery of a new protein family involved in cell membrane structure and function in diatoms.

The *P. multiseries* cultures used for RNA extraction were non-axenic. However, bacterial RNA contamination was expected to be a minimal concern in the *P. multiseries* EST database because most bacteria do not synthesize polyadenylated RNA during mRNA transcription. A number of findings support the view that the presence of bacteria in the *P. multiseries* cultures did not contribute to the content of the cDNA library. Sequence analysis of the *P. multiseries* cDNAs did not reveal any obvious contamination concerns that might have arisen from the presence of a poly (A)+ bacterial contaminant, as discussed below. In addition, mRNAs extracted from presumably axenic cultures that

were hybridized to the *P. multiseries* cDNA microarray reported in Chapter 3 of this thesis further confirmed the assignment of these mRNAs as *P. multiseries* transcripts. The library did not reveal rRNA fragments or other contaminating sequences, validating the overall quality of the poly (A)+ RNA used to construct this library.

The *P. multiseries* deduced amino acid sequences that matched most closely with bacterial proteins were searched against the *T. pseudonana* and *P. tricornutum* databases to evaluate if these were contaminating sequences from bacteria in the cultures (Table 2-4). The ten alignments with the highest E-values were chosen for this analysis. Each of these sequences matched most closely with the other diatom sequences, supporting the validity of the designation of these transcripts as derived from *P. multiseries*. Most of these sequences matched most closely to *P. tricornutum*, consistent with the closer evolutionary relationship between the two pennate diatoms (Kooistra et al., 2003; Damste et al., 2004).

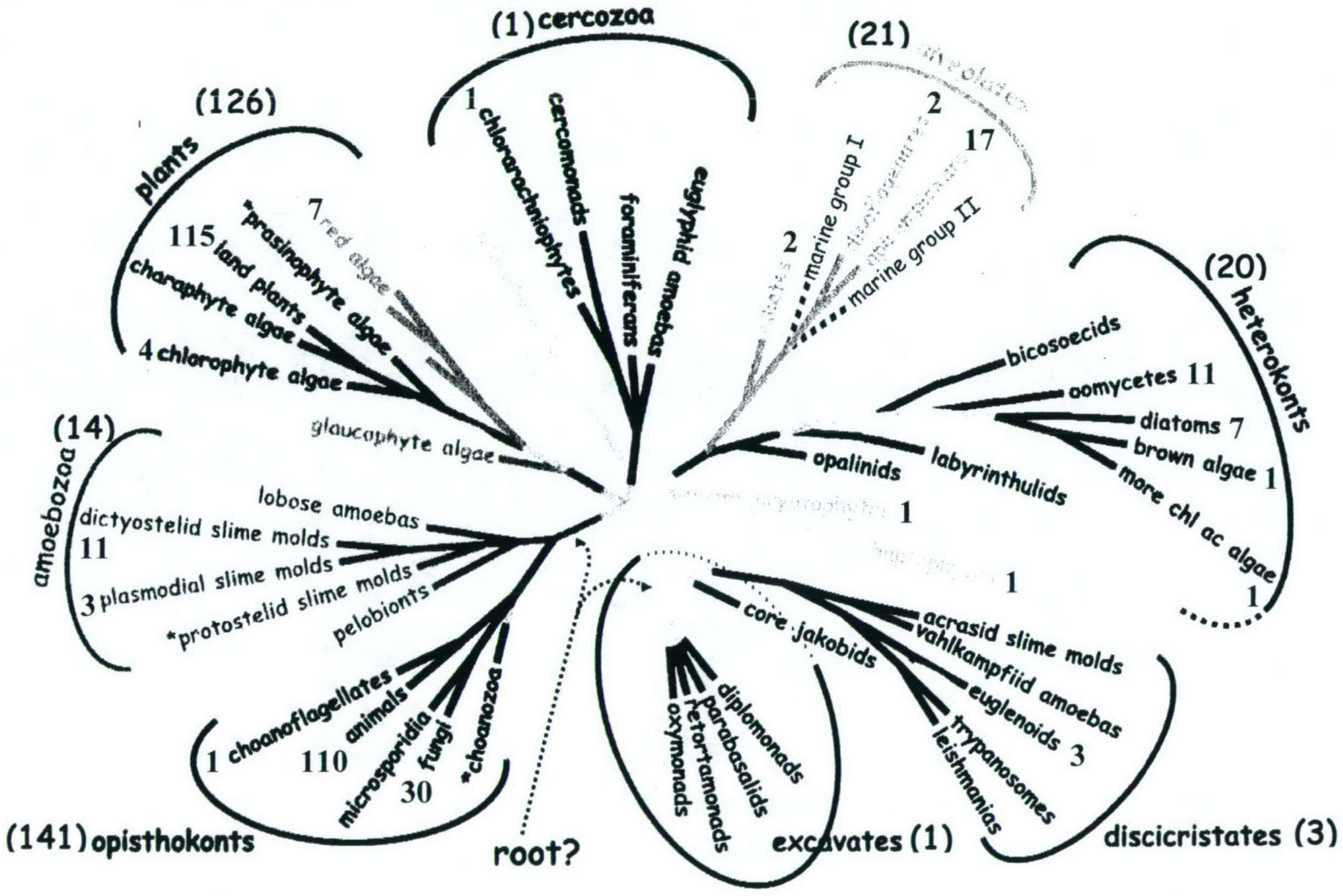
An emerging model of phylogenetic relationships among the eukarya, using combined data from rRNA, alpha-tubulin, beta-tubulin, actin, and elongation factor-1 alpha (EF-1 alpha) has revealed 8 major groups (Figure 2-5)(Baldauf et al., 2000, Baldauf, 2003). Multi-gene datasets for taxa within these groups are necessary to facilitate the resolution of the branches of the eukaryotic tree and to further define the root of the tree. Multiple copies of actin, beta-tubulin, and EF-1 alpha were identified in *P. multiseries*. These genes and others, such as the chaperone, Hsp70, should assist in reconstructing phylogenetic relationships both within the *Pseudo-nitzschia* spp., and within its major group, Heterokonta. Lundholm et al. (2002) suggest a paraphyletic origin of *Pseudo-nitzschia* spp., based on rRNA and morphological data. Multi-gene data sets are more reliable than single gene studies, and additional analyses using other genes are necessary to validate this conclusion and further define the relationship among toxin and non-toxin-producing strains of *Pseudo-nitzschia*.

Table 2-4. The Thalassiosira pseudonana and Phaeodactylum tricornutum databases were searched for sequences with high similarity to the *P. multiseries* ESTs that aligned most closely with bacterial proteins. The E-value, % identity, and % similarity are presented for the original Blast search against the non-redundant (NR) database, and % identity, % similarity are presented from the searches against the T. pseudonana and P. tricornutum databases.

	PSN I.D.	Putative Identification	*Species* or Domain Name	NR	NR	*Thalassiosira pseudonana*	*Phaeodactylum tricornutum*
1	PSN0016	Phosphoenolpyruvate carboxykinase	*Campylobacter jejuni*	1.E-164	55%, 67%	61%, 67%	73%, 85%
2	PSN0081	Acyl-CoA dehydrogenase	*Xanthomonas campestris*	1.E-117	47%, 63%	71%, 78%	80%, 86%
3	PSN0037	Long-chain acyl-CoA synthetases	*Pseudomonas fluorescens*	1.E-95	50%, 60%	61%, 72%	No Hit
4	PSN0100	PPi-phosphofructokinase	*Pirellula sp. 1*	4.E-67	53%, 69%	65%, 82%	68%, 82%
5	PSN0054	Long-chain acyl-CoA synthetases	*Rubrobacter xylanophilus*	1.E-66	35%, 49%	56%, 69%	65%, 81%
6	45H7	Acyl-coenzyme A synthetases	*Rhodobacter sphaeroides*	2.E-66	52%, 65%	73%, 82%	64%, 72%
7	PSN0095	Nucleotide sugar epimerase	*Gloeobacter violaceus*	2.E-60	40%, 55%	81%, 92%	No Hit
8	175A9	Phosphoglycerate mutase	*Chlorobium tepidum*	4.E-53	55%, 67%	74%, 85%	76%, 85%
9	186D10	Phosphoglycerate mutase	*Geobacter sulfurreducens*	3.E-51	48%, 63%	50%, 67%	76%, 83%
10	7B8	GTP cyclohydrolase II, putative	*Bradyrhizobium japonicum*	7.E-29	58%, 70%	66%, 69%	No Hit

** No Hits against the *P. tricornutum* database does not necessarily indicate the lack of a similar sequence, because the EST database is not exhaustive. However, the T. pseudonana database consists of the complete genome sequence for this diatom, so it was encouraging to discover highly similar sequences for each of these ESTS, and this probably indicates that these putatively identified proteins are also present in P. tricornutum.

Figure 2-5. A consensus phylogeny of eukaryotes, showing the number of *P. multiseries* sequences that showed significant sequence similarity to coding sequences from species with the given group. Modified from Baldauf et al., 2003.



As indicated on the eukaryotic tree, among the *P. multiseries* sequences showing significant similarity to known proteins, several were most similar to coding sequences in other diatoms. These included a silicon transporter, a fucoxanthin-chlorophyll a/c light-harvesting protein, two glyceraldehyde-3-phosphate dehydrogenases, a delta-6 fatty acid desaturase, a phosphoglycerate kinase precursor protein, and two chaperones, BiP and Hsp70. In addition, numerous significant similarities were found within the major heterkont group, including a 6-phosophogluconate dehydrogenase, an electron flavoprotein beta subunit, a GTP-binding protein, an S-adenosyl methionine synthetase, a histone H3, a number of different actin sequences, and the chaperones Hsp70 and Hsp 90-1. The multiple actin hits may represent a multiple-copy actin gene family, which has been demonstrated in the oomycetes, *Lagenidium giganteum* and *Pythium irregulare* (Bhattacharya and Stickel, 1994.)

The functional classification of derived coding sequences from *P. multiseries* revealed that proteins involved in translation represented a large proportion of the EST database (12.5%). The most abundant mRNA in the entire *P. multiseries* sequence assembly was EF-1 alpha, with 73 representative cDNAs (Table 2-5). Other cDNA libraries have also observed high representation of EF-1 alpha, such as the *L. digitata* library (Crepineau et al., 2000). EF-1 alpha modulates a diverse range of cellular activities, including protein synthesis, cell growth, motility, protein turnover, and signal transduction (Ridgely et al., 1996). The critical role of EF-1 alpha in regulating cellular activities suggests that it is essential to *P. multiseries* biology.

The *P. multiseries* deduced amino acid sequence was most similar to EF-1 alpha in the choanoflagellate, *Monosiga brevicollis*, yielding an e-value of 1E-115, with identity and similarity of 51% and 70%, respectively. Searching this *P. multiseries* sequence against the *T. pseudonana* genome detected five possible EF-1 family members. The best hit produced an alignment of 1303 bp, with identity and similarity of 83% and 87%, respectively. Blast analysis of the *T. pseudonana* sequence against the nr database also demonstrated the highest similarity to *Monosiga brevicollis*, with an E-value of 1E-120, identity, 49%, and similarity, 67%. While the *P. tricornutum* database also

Table 2-5. Most prevalent mRNAs as Measured by Redundancy

	PSN Contig I.D.	cDNAs per contig	Contig Length (bp)	NCBI Accession No.	Putative Identification	*Species* or Domain Name	E-value
1	PSN0001	73	1628	AAK27413.1	Elongation factor 1alpha long form	*Monosiga brevicollis*	e-115
2	PSN0002	53	2445	----------	Novel sequence, no significant hits**	----------	----------
3	PSN0007	42	1672	ZP_00161283.1	Uncharacterized protein with von Willebrand factor type A (vWA) domain	*Anabaena variabilis*	3.00E-07
4	PSN0006	28	1262	NP_520588.1	Transmembrane protein, probable	*Ralstonia solanacearum*	0.056
5	PSN0009	24	1241	NP_199564.1	Auxin-responsive protein, putative	Arabidopsis thaliana	0.004
6	PSN0011	22	2263	ZP_00064353.1	3-carboxymuconate cyclase	COG2706	8.00E-48
7	PSN0280	20	1261	----------	Novel sequence, no significant hits**	----------	----------
8	PSN272	18	1029	----------	Novel sequence, no significant hits**	----------	----------
9	PSN0013	16	1895	----------	Novel sequence, no significant hits**	----------	----------
10	PSN0031	15	1980	BAB86297.1	Alkaline serine protease IV	*Alteromonas sp. O-7*	2.00E-46
11	PSN0016	14	2158	NP_282084.1	Phosphoenolpyruvate carboxykinase (ATP)	*Campylobacter jejuni*	e-164
12	PSN0183	14	2921	ZP_00066068.1	Secreted protein containing C-terminal beta-propeller domain distantly related to WD-40 repeats	*Microbulbifer degradans*	2.00E-11
13	PSN0014	13	2438	NP_662047.1	Long-chain-fatty-acid-CoA ligase	*Chlorobium tepidum*	4.00E-48
14	PSN0169	13	1602	----------	No significant Hits**	----------	----------
15	PSN0759	10	1433	NP_011036.1	Swi4p, Involved in cell cycle dependent gene expression	*Saccharomyces cerevisiae*	0.006
16	PSN0273	10	1279	NP_867660.1	Protein-signal peptide and transmembrane prediction	*Pirellula sp. 1*	0.004

**Novel sequences did not show any sequence similarity to Thalassiosira pseudonana genome sequence, while PSN0169 showed 69% positives and 50% identity against T. pseudonana.

appeared to include numerous elongation factors, the highest similarity found to the putatively identified *P. multiseries* EF-1 alpha was 56% identity and 75 % similarity. The *P. tricornutum* sequence showed strong similarity to EF-1 alpha from *Euglena gracilis*, represented by an E-value of 1E-104, with 84% identity and 90% similarity. The sequence data available from these three diatom projects will allow further examination of the organization and expression of EF-1 alpha genes to determine both functionality and divergence of the EF-1 alpha genes within these groups.

The *P. multiseries* cDNA library also included four novel sequences that were highly expressed. These sequences did not match any sequences in the other diatom databases, nor the public nr protein and nucleotide databases. Therefore, further characterization of these transcripts may offer valuable insight into unique aspects of *P. multiseries* biology. Other highly redundant mRNAs included three transcripts that were up-regulated during toxin production; these included 3-carboxymuconate cyclase, phosphoenolpyruvate carboxykinase, and a long-chain fatty acid CoA ligase (discussed in the next chapter). In addition, another highly expressed mRNA showed high sequence similarity to a subtilisin-type alkaline serine protease. These peptidases may be involved in cell wall synthesis or scavenging nutrients from the environment, so this transcript may also reveal insight into either of these important activities in *P. multiseries* biology (Miyamoto et al., 2002; Siezen and Leunissen, 1997; Graycar, 1999).

Surprisingly, only one *P. multiseries* cDNA sequence coded for fucoxanthin, chlorophyll a,c-binding protein (FCP), and 2 cDNAs coded for other light harvesting proteins (LHP). The FCPs are major components of the photosystem II-associated light harvesting complex in diatoms and other brown algae (Bhaya and Grossman, 1993). In both the *L. digitata* and *P. tricornutum* EST databases, FCPs and LHPs were multigenic and represented highly redundant mRNAs (Crepineau et al., 2000; Scala et al., 2002). In the public nr protein database, *P. multiseries* FCP aligned most closely with the diatom *Skeletonema costatum* (E-value 2E-50, 63% identity, 69% similarity), which was also reported to contain multiple copies of this gene. Searching the *P. tricornutum* EST database revealed 63% identity and 76% similarity with one of the *P. tricornutum* FCPs.

This single member of the FCP multi-gene family was represented by 18 separate cDNAs in the *P. tricornutum* database. *P. multiseries* array experiments in the next chapter confirmed that *P. multiseries* FCP was down-regulated during toxin production. Leblanc et al. (1999) monitored FCP expression in dark-adapted cultures of the centric diatom *Thalassiosira weissflogii* and found that mRNA levels increased 5- to 6- fold in response to white light irradiation. In the growth experiments used for the cDNA library preparation, cells were grown at 100 $\mu E\ m^{-2}\ s^{-1}$, 14:10 h LD cycle. In the growth experiments completed for the differential expression studies, cells were grown at 100 $\mu E\ m^{-2}\ s^{-1}$, 24 L. Cells were harvested during the light cycle in both experiments. So, down-regulation does not appear to be induced by response to changes in light regime in the *P. multiseries* experiments. Oeltjens et al. (2004) showed that steady-state mRNA concentrations of FCP in the centric diatom *Cyclotella cryptica* oscillated in a circadian manner. Again, the differences in culture conditions and harvesting during the light cycle would suggest that circadian rhythms were not controlling FCP expression in *P. multiseries*. However, down-regulation of FCP in *P. multiseries* was correlated with stationary growth, when photosynthesis would presumably decrease as cell growth slows due to some limiting factor. The pathways leading to chlorophyll and DA production may both draw on a pool of glutamate (Bates et al., 1998), therefore, the down-regulation of FCP also correlates well with the onset of DA production. The *P. multiseries* FCP sequence identified in this study can now be used to probe for nuclear-encoded FCPs of this gene family in *P. multiseries*, and to further investigate FCP regulation and control in *P. multiseries*.

The *P. multiseries* EST database also led to the discovery of a protein coding sequence demonstrating high similarity to an enzyme involved in the C4 pathway of photosynthetic carbon assimilation. This transcript shared 67% similarity, and 53% identity (E-value E-140) with a C4-specific pyruvate, orthophosphate dikinase (PPDK) from *Miscanthus x giganteus* (Naidu et al., 2003). PPDK is localized to chloroplasts in C4 plants and catalyzes the conversion of pyruvate to phosphenolpyruvate. An amino-terminal sequence of the C4-PPDK directs entry of the precursor protein into chloroplasts

(Agarie et al., 1997). The separation of enzymes involved in C4 and Calvin cycles into cellular compartments would allow C4 photosynthesis to occur in a single-celled organism, such as *P. multiseries*, without the complex tissue structure of higher plants. Other key enzymes involved in C4 photosynthesis that were found in the *P. multiseries* EST database included phosphoenolpyruvate carboxykinase, phosphoenolpyruvate carboxylase, and pyruvate carboxylase. The existence of a C4 photosynthetic pathway in diatoms has been debated (Reinfelder et al., 2000; Johnston et al., 2001), and the discovery of a potential C4-specific PPDK in *P. multiseries* suggests the exciting possibility that a C4 mechanism is active in *P. multiseries*. This discovery would potentially contribute to the revision of current hypotheses on the evolutionary history of C4 photosynthesis and provide further insight into the photosynthetic activities and ecological success of marine diatoms. (This hypothesis is discussed further in chapter 4.)

Ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) is a another principal carbon fixation enzyme, which is alleged to represent the most abundant enzyme on earth (Barraclough, 1979; Smith, 1981). While mRNAs encoding this enzyme have been found in abundance in plant EST databases (ex. Hofte et al., 1993), diatoms are known to have plastid-encoded rubisco, which would account for why this gene was not identified in the *P. multiseries* library (Hwang and Tabita, 1989, 1991) These results are consistent with those found in the *P. tricornutum* EST database.

The *P. multiseries* database included a high number of fatty acid and lipid molecules, which may be involved in cell membrane synthesis, fuel for metabolism, or synthesis of the DA isoprenoid side-chain. In addition, many lipid molecules mediate signal transduction. Enzymes that control production of lipid signaling molecules in plants include phospholipases, lipid kinases, and phosphatases (Wang, 2004). Diatoms must respond to constantly changing environmental conditions, so signal transduction pathways are important to their survival. In addition to numerous other signaling molecules found in *P. multiseries*, three potential lipid-signaling enzymes were identified, including inositol 5-phosphatase, and two phospholipases. One of these enzymes, phospholipase A2, appears to activate defense response in the diatom,

Thalassiosira rotula (Pohnert, 2002). The study of lipid signaling in plants is still in its early stages, so the discovery of a number of genes that are potentially involved in lipid signaling pathways may offer an opportunity to facilitate the advancement of our understanding of these pathways in *P. multiseries* and other photosynthetic organisms.

Other *P. multiseries* transcripts of interest included one likely to encode a silicon transporter, SIT. Silicon transport is essential to silica metabolism, so the identification of SIT offers a useful tool to study cell wall synthesis in *P. multiseries* (Hildebrand et al., 1998). *P. multiseries* cDNAs with significant similarity to ferredoxin and flavodoxin coding sequences may prove useful for exploring iron limitation in *Pseudo-nitzschia* spp. (McKay, 1997; Erdner et al., 1999). Finally, *P. multiseries* sequences that appear to encode cell division genes, such as cell division cycle 27, may facilitate the development of new methods for measuring populations growth rates in *Pseudo-nitzschia* spp. (Lin et al., 1998, 1999, 2000).

Discussion:

The EST database provides original information on the expressed genome of *P. multiseries*, which will help to facilitate further studies into the physiology, ecology, and evolutionary history of this organism. Comparative studies among the three diatoms, *T. pseudonana*, *P. tricornutum*, and *P. multiseries*, will likely facilitate further understanding of the intricacies of diatom biology through molecular genomics. This work represents an entry into the study of metabolic pathways in *P. multiseries*, and has begun to reveal new information about *P. multiseries* biology. For example, the presence of novel sequences that did not show sequence similarity to any of the sequences in the *T. pseudonana* or *P. tricornutum* databases suggests that this diatom contains divergent sequences that are specific to the biology of *P. multiseries*.

The genome size of *P. tricornutum*was recently estimated to be 13 Mb (± 6 Mb) (Scala et al, 2002). *T. pseudonana* genome size has been estimated to be 34.3 Mb, while the number of protein encoding genes in *T. pseudonana*has been estimated to be approximately 11,000 protein genes (http://genome.jgi-psf.org). It is likely that a genome size for *P. multiseries* is in the same range of these diatoms. The estimate of EST number identified in our study of *P. multiseries* (~4,000) under the specific physiological states is lower than the 11,000 described for *T. pseudonana*. However, it is reasonable to presume that additional transcripts will be discovered through the additional characterization of the current cDNA library as well as the the study of cDNAs derived from additional physiological states.

As the sequences that are novel to *P. multiseries* are further characterized, they may offer a useful tool for looking at evolutionary relationships within the *Pseudo-nitzschia* spp. Genes associated with toxin production will be most useful for understanding the relationships between toxin- and non-toxin-producing *Pseudo-nitzschia* spp., and for monitoring toxin production in the field. Many possible candidate genes that may play a role in DA biosynthesis were revealed in the EST sequencing project. Examples include genes likely to encode enzymes involved in isoprenoid,

pyruvate, or glutamate metabolism, such as delta 6 fatty acid desaturase, phosphoenol-pyruvate carboxykinase, glutamate dehydrogenase, and 5-oxo-L-prolinase. cDNA array experiments were designed in the next chapter to select for genes that were specifically correlated with toxin production; this dataset offers useful target genes for further characterization, which should lead to a better understanding of *P. multiseries* biology and DA biosynthesis, both in the lab and field.

The EST study contributes 411 newly identified coding sequences from *Pseudo-nitzschia multiseries*. This data can now be used to identify nuclear-encoded genes from *P. multiseries* or other related diatoms and to further characterize the role of specific genes in *P. multiseries* biology.

Chapter III

Gene Expression Profiling

Abstract:

A cDNA microarray was designed to screen for differentially expressed genes under toxin-producing vs. non-toxin-producing conditions in *Pseudo-nitzschia multiseries*, in order to begin to understand the biochemical pathways and physiological control mechanisms which relate to toxin production in this organism. Expression analysis of 5,372 cDNAs revealed 121 up-regulated cDNAs, representing 12 unique transcripts, and 51 down-regulated cDNAs, representing 15 unique transcripts. Up-regulated transcripts encoded protein sequences with structural similarity to a 3-carboxymuconate cyclase, phosphoenolpyruvate carboxykinase, an amino acid transporter, a small heat shock protein, a long-chain fatty-acid-CoA ligase, and an aldo/keto reductase. Down-regulated transcripts included sequences with similarity to a key regulatory enzyme involved in glycolysis, Ppi-phosphofructokinase, and a light harvesting protein, fucoxanthin-chlorophyll a/c light harvesting protein. These results provide a framework for investigating the control of toxin production in *P. multiseries*. These transcripts may also be useful in ecological field studies in which they may serve as signatures of toxin production.

Introduction:

Domoic acid (DA) is a phycotoxin produced by a group of marine algae limited to certain species of the diatom genera *Pseudo-nitzschia, Nitzschia*, and *Amphora,* and the macro red algae *Chondria, Alsidium, Amansia, Digenea, and Vidalia* (Takemoto and Daigo, 1958; Wright et al., 1989; Bates et al., 1998; Bates, 2000). Accumulation of DA by filter feeding of *Pseudo-nitzschia* cells and subsequent transmission of the neurotoxin to humans via shellfish has resulted in severe illness, designated amnesic shellfish poisoning (ASP) due to symptoms characterized by memory loss (Bates et al., 1989, Bates, 1998, Wright et al., 1989). DA is a neuroexcitatory amino acid that exhibits structural similarity with glutamic acid, kainic acid, and proline (Figure 1-1). DA binds to glutamic acid receptors with an affinity up to 100 times that of glutamate, leading to prolonged depolarization and ultimately swelling and cell death in neurons exposed to this water soluble amino acid (Stewart et al., 1990; Olney, 1994). Efforts to discover the environmental factors that stimulate DA production by *Pseudo-nitzschia* spp. have led to a greater understanding of the physiology and ecology of these organisms, yet the characterization of the biosynthetic pathways leading to DAsynthesis has been minimal, limited to two ^{13}C- and ^{14}C-labelling studies (Douglas et al., 1992; Ramsey et al., 1998) and more recently to a computational modeling approach (Smith et al., 2001; Smith, personal communication).

The carbon labeling experiments supported condensation of an activated glutamate derivative from the citric acid cycle with an isoprenoid chain, such as geranyl pyrophosphate, and subsequent cyclization as a possible mechanism for DA biosynthesis (Figure 1-2). On the other hand, Smith and colleagues have focused on the relationship of proline to DA metabolism, by modeling and measuring amino acid levels to show that proline and DA are inversely correlated, therefore, suggesting that either 1) proline is an upstream precursor to DA, or 2) DA substitutes for the physiological function of proline. Smith goes on to suggest a biochemical model describing the hypothesized derivation of 3-hydroxy-glutamate from proline metabolism, leading to DAsynthesis (Figure 1-3).

The two proposed models for DA synthesis are linked by 3-hydroxy-glutamate, which the former proposal extends to suggest condensation of this glutamate derivative with an isoprenoid chain and subsequent cyclization to form the pyrrolidine ring of DA. Both of these proposed pathways suggest many potential metabolic schemes, and further understanding of DA biosynthesis would be limited without an investigation into the genes that govern the regulation of these pathways. Therefore, the goal of this study was to identify genes that are up-regulated during toxin production in an effort to advance our understanding of DA biosynthesis and regulation, and to provide further insight into the overall physiology of *P. multiseries.*

DA production has been shown to begin during the late exponential growth phase and peak during the stationary phase, when division of the entire population of cells slows due to Si or P limitation (Bates, 1998). This study applied cDNA microarray technology to investigate gene expression in *P. multiseries* during high-toxin-producing vs. low-toxin-producing conditions by comparing mRNAs from cells that were in exponential phase to cells that were in stationary phase. Comparative analysis of cells harvested over the growth cycle of *P. multiseries* would likely select for genes associated with DA biosynthesis, transport, cell cycle progression, cell signaling, reproduction, and stress response. The construction of *the P. multiseries* cDNA library (chapter II) facilitated the manufacture of *P. multiseries* cDNA microarrays to screen the library for genes or clusters of genes that were up-regulated during toxin production. Analysis of this large set of expression data has revealed several candidate genes that may be involved in DA biosynthesis, stress response, and carbohydrate metabolism.

Materials and Methods:

The technical and software options in microarray analysis are vast and no single protocol is available for individual cDNA array projects; therefore, protocols must be optimized for each individual application of this technology. Alternative protocols were evaluated throughout each step of the cDNA microarray approach to expression profiling in *Pseudo-nitzschia multiseries*, from array construction through data analysis. The final protocols used in the *P. multiseries* project are presented in the following sections, with notes on alternatives, when useful or informative.

Growth Experiments: Pseudo-nitzschia multiseries strains used in this study were graciously provided by Stephen S. Bates (Department of Fisheries and Oceans, Gulf Fisheries Center, Moncton, NB, Canada.) The strains included CLN-125, CLN-125 – Axenic, and CLN-191. *P. multiseries* cells were grown in 0.2µm filtered seawater enriched with f/2 nutrients (Guillard, 1975). Initial inoculum was acclimated to experimental culture conditions, and cells were in exponential growth phase. Batch cultures were maintained at 20°C, 100µEm^{-2}s^{-1}, 24 h Light. Fifteen L of culture were grown in 19-L borosilicate carboys; cultures were aerated using an aquarium pump and sterile tubing and the cultures were constantly mixed with magnetic stirrers.

Samples were taken every two to three days for cell counts, DA analysis, and nutrient analysis. Cell concentrations were estimated by averaging the number of cells enumerated by light microscopy using a Neubauer hemacytometer chamber in three separate counts of individual samples preserved in Lugol's iodine. DA concentrations were analyzed in whole culture samples (cells plus medium) by Claude Leger in Stephen S. Bates' laboratory using a FMOC derivatization method (Bates et al., 1989, Pocklington et al., 1990).

From the original 15 L of *P. multiseries* culture grown per carboy, eight L of culture were harvested at an initial time point during mid- to late exponential growth (Harvest 1). The remaining seven L of culture were harvested at a final time point during

Table 3-1: DA Concentrations for *Pseudo-nitzschia multiseries* Growth Experiments: CLN-125 Axenic, Experiments 1-4; CLN-125, Experiments A-D; CLN-191, Experiments A-D. Experiments in red represent those used for microarray studies.

EXPERIMENT - Harvest	DOMOIC ACID (ng/ml)
CLN 125 Axenic #1 -Day 9, Harvest 1	0
CLN 125 Axenic #1 -Day 42, Harvest 2	38
CLN 125 Axenic #2 -Day 9, Harvest 1	12
CLN 125 Axenic #2 -Day 42, Harvest 2	55
CLN 125 Axenic #3 -Day 10, Harvest 1	16
CLN 125 Axenic #3 -Day 30, Harvest 2	73
CLN 125 Axenic #4 -Day 10, Harvest 1	12
CLN 125 Axenic #4 -Day 30, Harvest 2	58
CLN 125A -Day 7, Harvest 1	18
CLN 125A -Day 9, Harvest 2	123
CLN 125A -Day 31, Harvest 3	1878
CLN 125B -Day 7, Harvest 1	24
CLN 125B -Day 9, Harvest 2	139
CLN 125B -Day 31, Harvest 3	2112
CLN 125C -Day 4, Harvest 1	2
CLN 125C -Day 10, Harvest 2	267
CLN 125D -Day 4, Harvest 1	0
CLN 125D -Day 10, Harvest 2	314
CLN 191A -Day 7, Harvest 1	547
CLN 191A -Day 31, Harvest 2	10031
CLN 191B -Day 7, Harvest 1	617
CLN 191B -Day 31, Harvest 2	7768
CLN 191C -Day 2, Harvest 1	44
CLN 191C -Day 8, Harvest 2	1218
CLN 191D -Day 2, Harvest 1	47
CLN 191D -Day 8, Harvest 2	1319

stationary growth (Harvest 2). The cell suspension was spun in 0.5 L bottles for 15 minutes at 1000*g*. The resultant pellets were pooled, split among 2-4, 50ml conical tubes and spun briefly to remove any remaining liquid. Ten to 20 mL of Trizol were added to the conical tubes, and the pellets were homogenized for 60 seconds at full speed (Polytron), frozen in liquid N, and stored at -80°C for later RNA extraction. The goal was to harvest cells during high (stationary growth) vs. low (exponential growth) DA producing conditions. DA analysis revealed that three out of twelve growth experiments had undetectable or minimal DAconcentrations at the initial harvest and relatively high DAconcentrations at the final harvest (Table 3-1). Therefore, these three experiments were selected for further analysis using the *P. multiseries* cDNA microarrays. The three growth experiments were designated 125C, 125D, and AX1, referring to CLN-125 (non-axenic), growth experiments C and D, and CLN-125 (Axenic), growth experiment 1, respectively.

Construction of P. multiseries cDNA microarray: A total of 5372 clones from the *P. multiseries* cDNA library were grown overnight in Luria broth with carbenicillin (50 µg /ml), at 37°C on a shaker table. A volume of 10 µL of bacterial culture was then used as template in 100 µL PCR reactions with primers T7 forward (TAATACGACTCACTA TAGGG) and M13 reverse (CAGGAAACAGCT ATGAC), which flank the cloning site of the pMD1 (a pUC18-derived) vector (see library construction, chapter 2). PCR conditions were optimized to include the following reagents: 1X PCR buffer (Invitrogen), 200µM each dNTP, 2µM each primer, 2mM MgSO4, and 2.5U Invitrogen HiFI Taq polymerase. An initial DNA denaturation step at 94°C for 2 minutes was followed by 35 amplification cycles (0:30 melting at 94°C, 0:30 annealing at 55°C, 1:00 extension at 68°C). Samples of the bacterial clones used in PCR preparation were placed at -80°C in 15-30% glycerol, as back-ups of the original library clones.

PCR products were purified using Millipore MultiScreen size-exclusion filter plates. Vacuum pressure (approximately 10 inches Hg) was applied for 20 min or until wells were empty, to remove primers, dNTPS, and salts, while retaining the amplified DNA on the filter. A wash step included the addition of 50 to 100 µL of nuclease-free de-ionized water to each well; DNA was resuspended and mixed by repetitive pipetting, and the vacuum was re-applied. The DNA was then resuspended in 100 µL nuclease-free de-ionized water and transferred to clean plates using a mechanical pipetting station. The DNA was split into two aliquots; one for array printing and one for quality control and sequencing. DNA quality was verified by 1% agarose gel electrophoresis. DNA concentration was determined by PicoGreen fluorescent staining (Ahn et al., 1996). A limited number of samples were also quantified by measuring absorbency at 260/280nm to verify PicoGreen results. The final DNA concentrations averaged approximately 120 ng/µL. PCR product for printing was dried by vacuum centrifugation and resuspended in 10 µL of 1.5M Betain /3X SSC print buffer, yielding a final concentration of 600 ng/µL, on average.

P. multiseries cDNA probes were printed onto CMT-GAPS slides (Corning) using a Biorobotics MicroGrid 610 TAS Arrayer with quill pins. 5372 *P. multiseries* cDNAs were printed in duplicate; in addition, 10 control cDNAs from SpotReport Alien Array Validation System were printed in duplicate, resulting in a final chip including 10772 features. Spots were printed with a 32 print-tip head, producing a lay-out represented by 8 x 4 grids (Figure 3-1). Each grid was sub-divided into two sections, representing replicate spots (Figure 3-2). Individual features were 13 µm in diameter and were separated by 130 µm (from one spot to the next.) Approximately 0.005 µl of 600 ng/µL DNA (2-3ng) was transferred to each spot. Final *P. multiseries* arrays displayed strong signal to noise ratio, with virtually no background, as demonstrated visually (Figures 3-1 and 3-2). Results also illustrate the high degree of reproducibility between replicate spots on the *P. multiseries* chip.

Figure 3-1: Scanning Fluorescence Image of the *P. multiseries* cDNA microarray hybridized with Cyanine3 and Cyanine5 labeled cDNA from growth experiments. This image shows the microarray image, scanned at 595nm. 5376 individual *P. multiseries* cDNAs (X2) and 10 control cDNAs (X2) are represented on the array for a final chip including 10772 features. Note the uniform feature morphology and strong signal to noise ratio.

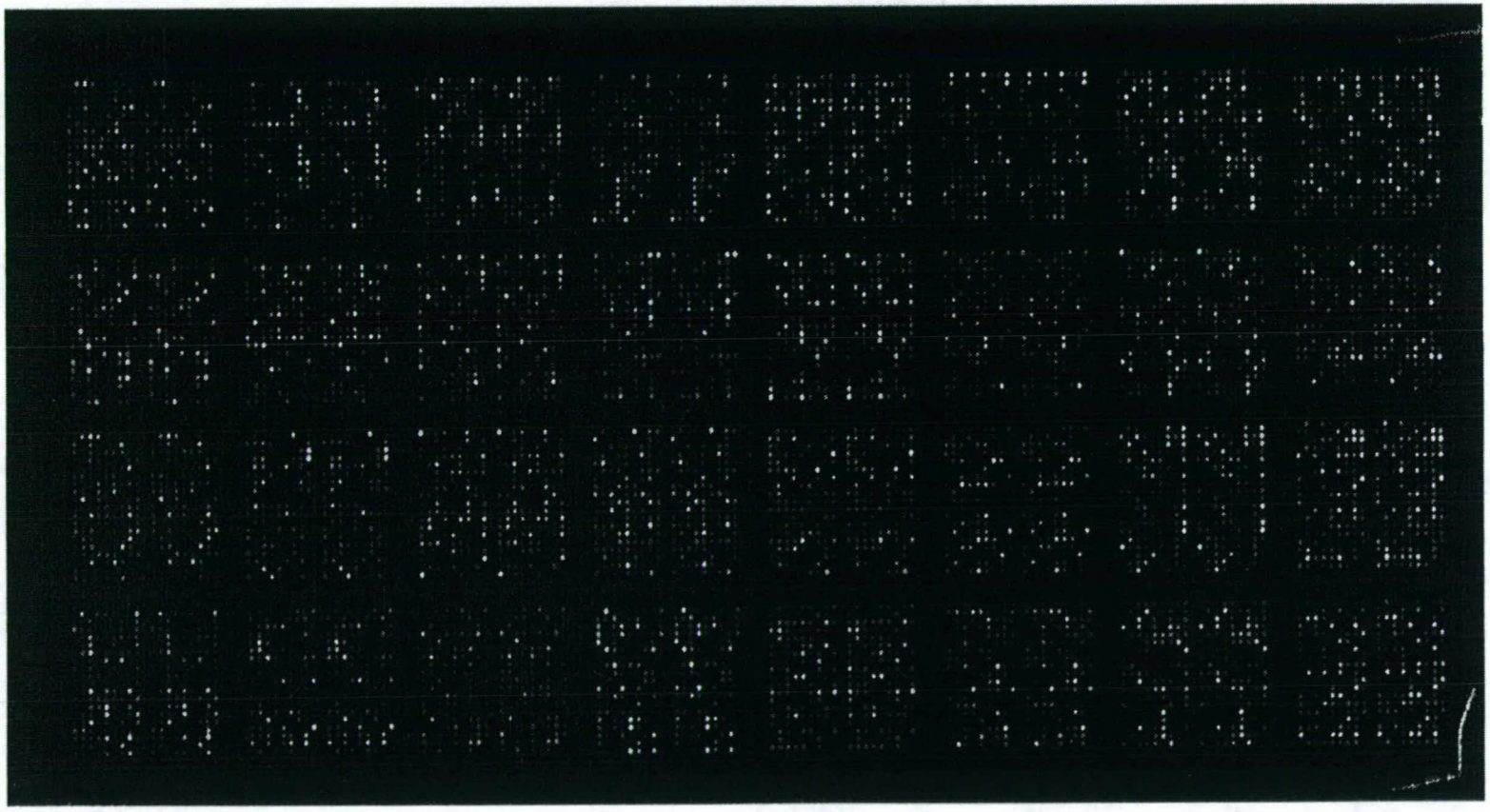
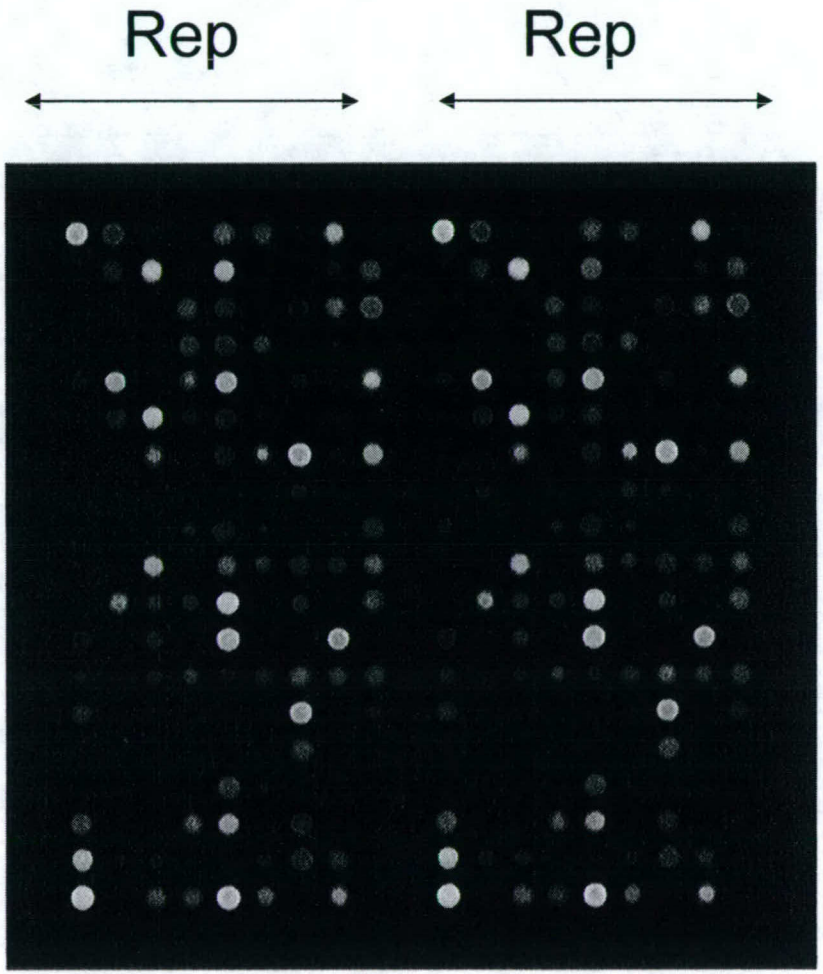


Figure 3-2: A representative grid from the Final *P. multiseries* cDNA microarray enlarged to demonstrate the high degree of reproducibility between replicate spots on the *P. multiseries* chip



Test print: A limited number of arrays were printed with 960 cDNA probes generated from the original *P. multiseries* cDNA library. The test print led to optimal protocols employed in the final print, presented above. During preparation of the test print, filter purification of PCR product using Millipore MultiScreen –PCR plates was compared to Qiaquick gel purification. Quality control of the resultant product by gel electrophoresis, PicoGreen quantification, and sequencing of resultant product showed that the methods were relatively comparable in quality, but filter purification resulted in much higher efficiencies of product recovery, by at least a factor of 10. A comparison of sequence length and quality in 96 samples yielded average reads of 616 bp for the Millipore filtered PCR product versus 513 bp for the Qiagen cleaned product. In addition, the filter screens were more time, labor, and cost efficient. Therefore, Millipore Multiscreens were adopted for subsequent purification of PCR product. Other trials using the test chip also allowed poly (A)+ hybridization to be compared to total RNA hybridization against the probe cDNAs on the microarray. Results illustrated that total RNA could be used successfully, without loss of signal or increased background. Finally, the total amount of RNA needed for target cDNA preparation was investigated, which led to a final protocol requiring 10μg of RNA per labeling reaction. This quantity of RNA was five-fold less than the original protocol required, which was helpful in experimental design and execution.

RNA Preparation and Microarray Hybridizations: Total RNA was extracted from *P. multiseries* cells harvested from growth Experiments 125C, 125D, and AX1 during high vs. low DA producing conditions (following RNA extraction procedure described in chapter 2.) Total *P. multiseries* RNA was cleaned with Qiagen RNeasy columns and run on formaldehyde agarose gels for quality control. (Gels were transferred onto Hybond membrane and stored in air-tight plastic bags at -20 °C for future analysis.) Ten micrograms *P. multiseries* RNA was spiked with mRNA from the Spot Alien Validation System, incubated for 10 minutes at 65°C with oligo-dT and then cooled at 25°C. Four μLs of 1mM Cy3- or Cy5- conjugated dUTPs were added and the mixture was incubated

at 42°C for two minutes. A master mix including 4.5µl 0.2M DTT, 18ul 5X 1st strand buffer, 1.8ul 25mM dATP, dGTP, and dCTP, and 1.8ul 10mM dTTP, and 2µl of Superscript II reverse transcriptase was added to the RNA mixture and incubated for 1 hour at 42°C. After one hour, an additional 1µl of Superscript II was added and the reaction was incubated at 42 °C for another hour. Starting RNA was degraded by addition of stop solution (3µl 0.5M EDTA, pH 8; 3µl 1N NaOH) and incubated for 30 min. at 60 °C. Labeled cDNA was cleaned up using Qiagen columns; Cy3 labeled cDNA and the corresponding Cy5 labeled cDNA that were to be compared were combined and loaded onto the same column. The labeled target cDNA pools were then hybridized to the probe cDNAs on *P. multiseries* microarrays.

Arrays were processed before hybridization as follows: the slides were humidified by holding them face-down over a steaming water bath for a few seconds, then snap-dried on a 95°C heat block. The DNA was immobilized onto the slides by UV cross-linking at 65mJoules. Cross-linked slides were soaked for 15 minutes in freshly prepared succinic anhydride/sodium borate solution with gentle agitation, soaked for 2 minutes in boiling nuclease free, de-ionized water and finally, rinsed in 95% ethanol and spun dry. Arrays were stored in a room temperature dessication chamber until hybridization.

Processed microarrays were pre-hybridized at room temperature for 1 hour. Pre-hybridization solution was composed of 50% formamide, 5X SSC, 0.1% SDS, 1% BSA, while hybridization buffer was composed of 50% formamide, 10X SSC, 0.2% SDS, 0.26% salmon sperm. Labeled cDNA was denatured prior to hybridization by heating for 2 minutes at 80°C, while the cassette and microarray were pre-warmed at 42°C. The cDNA was then loaded onto the array, and arrays were hybridized for 16 hours at 42°C in humidified chambers. The slides were then washed successively in 1X SSC, 0.03% SDS; 0.1X SSC, 0.01% SDS; and 0.1X SSC. Finally, the slides were dried by a brief centrifugation.

Experiments were dye-swapped to account for differences in dye labeling and detection efficiencies, for example, due to faster bleaching of Cy5 than Cy3. So, for each gene expression comparison, two hybridizations were completed with labeling of RNAs

being exchanged between the two dye-swap experiments. Technical replicates were also repeated within each experiment; for example, if 2 replicates were run, and two dye-swap experiments were carried out for each replicate, then there would be a total of 4 replicates to examine. Experiments 125C and 125D included a total of 6 replicate experiments, while AX1 included a total of 4 replicates.

Image analysis: Arrays were scanned at 595nm (Cy3) and 685nm (Cy5) on ArrayWoRx scanners (Applied Precision, Inc.) The ArrayWoRx scanning system converts signal from fluors to "pixel" values which allows the data to be saved as tiff files. MolecularWare DigitalGenome software was then used to integrate annotated chip information with the tiff files and to visualize, edit (ex. flagging spots covered by dust particles, missing spots, spots with low intensity, etc. for deletion), and export the data for further analysis. Data was exported into Microsoft Excel and sorted by Cy3 and Cy5 intensities to remove any data that was below an intensity level of 50 in both channels; the data was then normalized and analyzed for statistical significance.

Data normalization: Many sources of systematic variation may exist in microarray experiments that must be accounted for before expression levels can be compared appropriately. In this study, loess normalization was used to correct for differences in dye labeling and detection efficiencies, and other systematic biases in the measured expression levels both within and across arrays (Quackenbush, 2002; Park et al., 2003). The loess method of normalization scales individual intensities by fitting a curve to the data using a locally weighted non-linear regression, where $M = \log_2(Cy5/Cy3)$ for each element on the array is plotted as a function of $A = \log_{10}(Cy5*Cy3)$ product intensities. In these experiments, a loess algorithm was applied within and across each dataset using Insightful S+ArrayAnalyzer software (Figures 3-3 to 3-8). MvA and box plots for each of the *P. multiseries* growth experiments illustrate the normalization of data across the replicate arrays. Especially notable is the correction of Cy3 vs. Cy5 intensity differentials, illustrated by the fitting of the $\log_2(Cy5/Cy3)$ ratios to the average in each

of the box plots. (Other normalization parameters and options were considered, however, loess normalization yielded the most robust results, without washing out expression signals.)

Quality control included analyzing Cy3/Cy5 ratios for the control set of data after normalization. The normalized intensity data for each control spot was analyzed using linear regression analysis to verify that the total integrated intensity across the control spots was equal for both channels (slope = 1). The slope of the Cy3 to Cy5 linear regression approached 1 for all three experiments; AX1 slope was 0.98, while 125C was 1.17, and 125D was 0.96 (Figures 3-9 to 3-11). The variability around the slope of the Cy3/Cy5 ratio was especially small for Experiment AX1, and relatively small for Experiments 125C and 125D. Cy3/Cy5 ratios calculated individually for each feature in the whole dataset averaged 0.93 ± 0.09 in AX1, 0.98 ± 0.17 in 125C, and 1.11 ± 0.18 in 125D. In general, the standard deviation among replicate features in AX1 was less than in 125C and 125D. Therefore, in the statistical analysis that follows, more genes were called statistically significant in this experiment than in the other two experiments.

Statistical Analysis: Significance analysis of gene expression ratios was performed using a t-test algorithm modified for microarray analysis (Tusher et al., 2001). This method, Significance Analysis of Microarrays (SAM), identifies genes with statistically significant changes in expression by assimilating a set of gene-specific t-tests. SAM assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. A scatter plot of the observed relative difference d(i) vs. the expected relative difference dE(i) is used to identify significant changes in gene expression. For the majority of genes, d(i) approximates dE(i), but some genes are represented by points displaced from the d(i) = dE(i) line by a distance greater than a designated threshold, delta. Genes that fall outside the cutoff represented by delta are considered significant (figure 3-12 to 3-14). SAM generates a test statistic "q", which is similar to a p-value, but adapted to the analysis of a large number of genes. The q-value measures the significance of the expression ratio of a gene by reporting the lowest

Figure 3-3: Normalization of AX1 array data (Replicates represent dye swap experiments):

MvA plot illustrates the normalization of data across the replicate arrays. $M = \log_2(Cy5/Cy3)$. $A = \log_{10}(Cy5*Cy3)$.

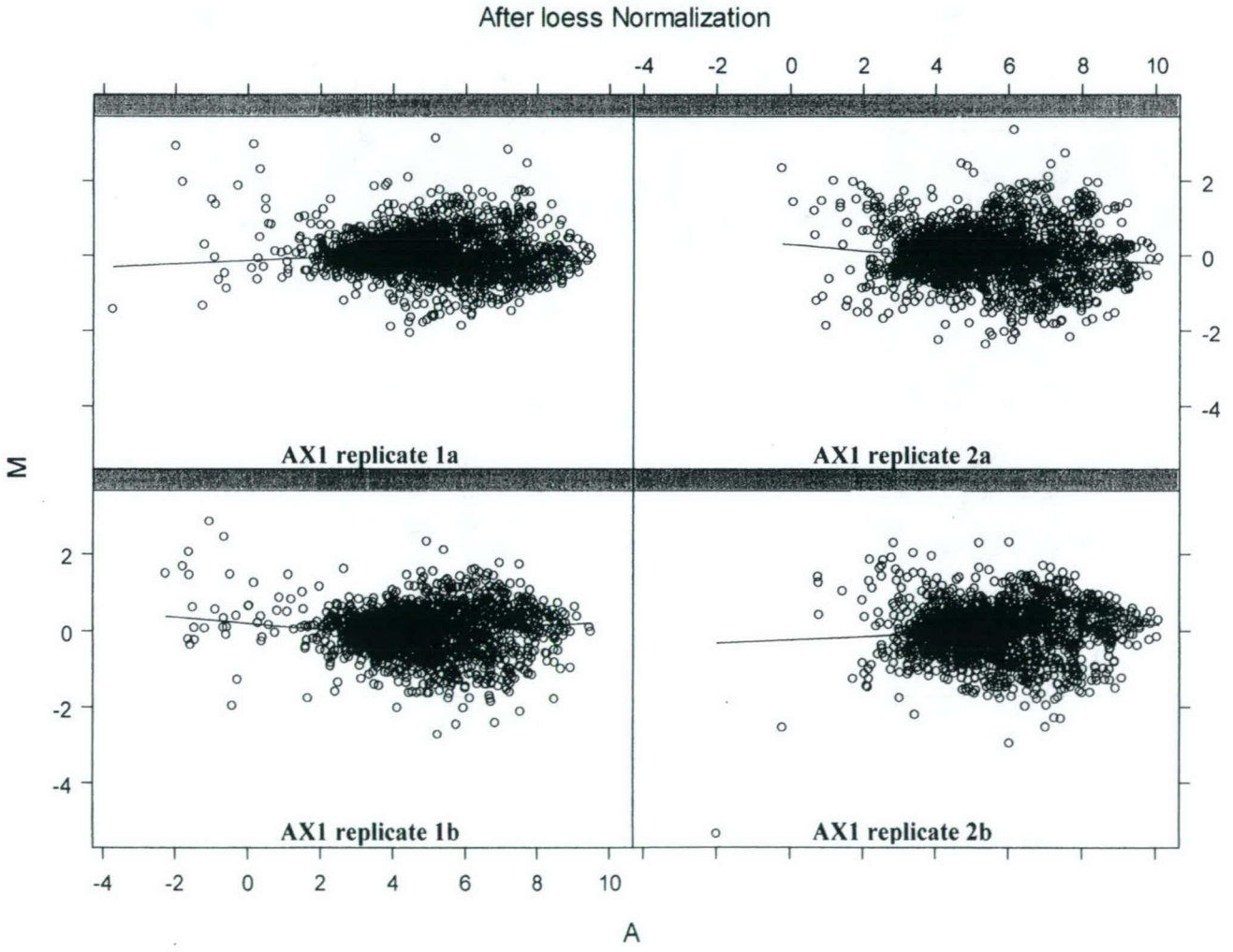


Figure 3-4: Normalization of AX1 array data (Replicates represent dye swap experiments):

Box plot illustrates the normalization of data across the replicate arrays. $M = \log_2(Cy5/Cy3)$.

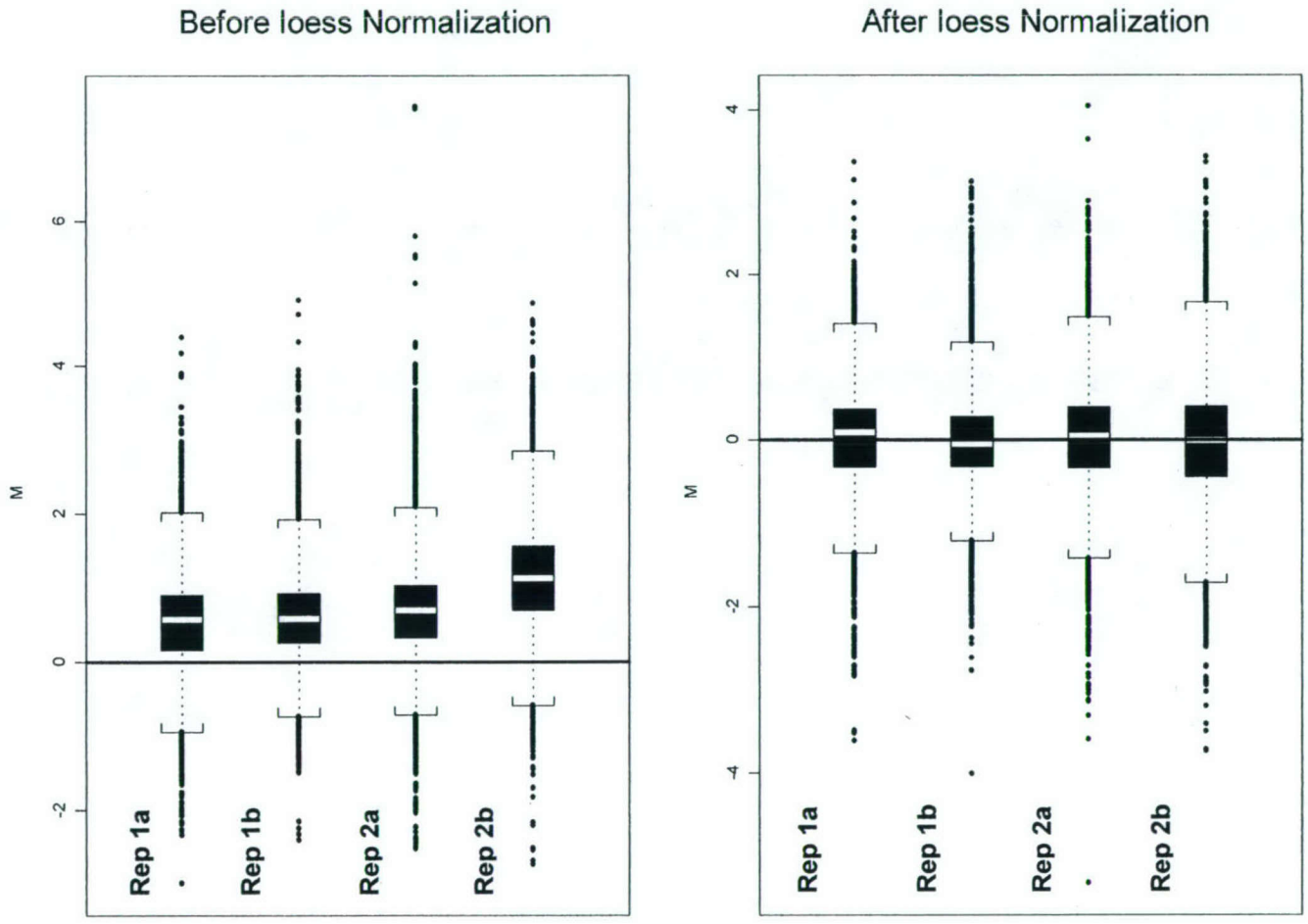


Figure 3-5: Normalization of 125C array data (Replicates represent dye swap experiments):

MvA plot illustrates the normalization of data across the replicate arrays. $M = \log_2(Cy5/Cy3)$. $A = \log_{10}(Cy5*Cy3)$.

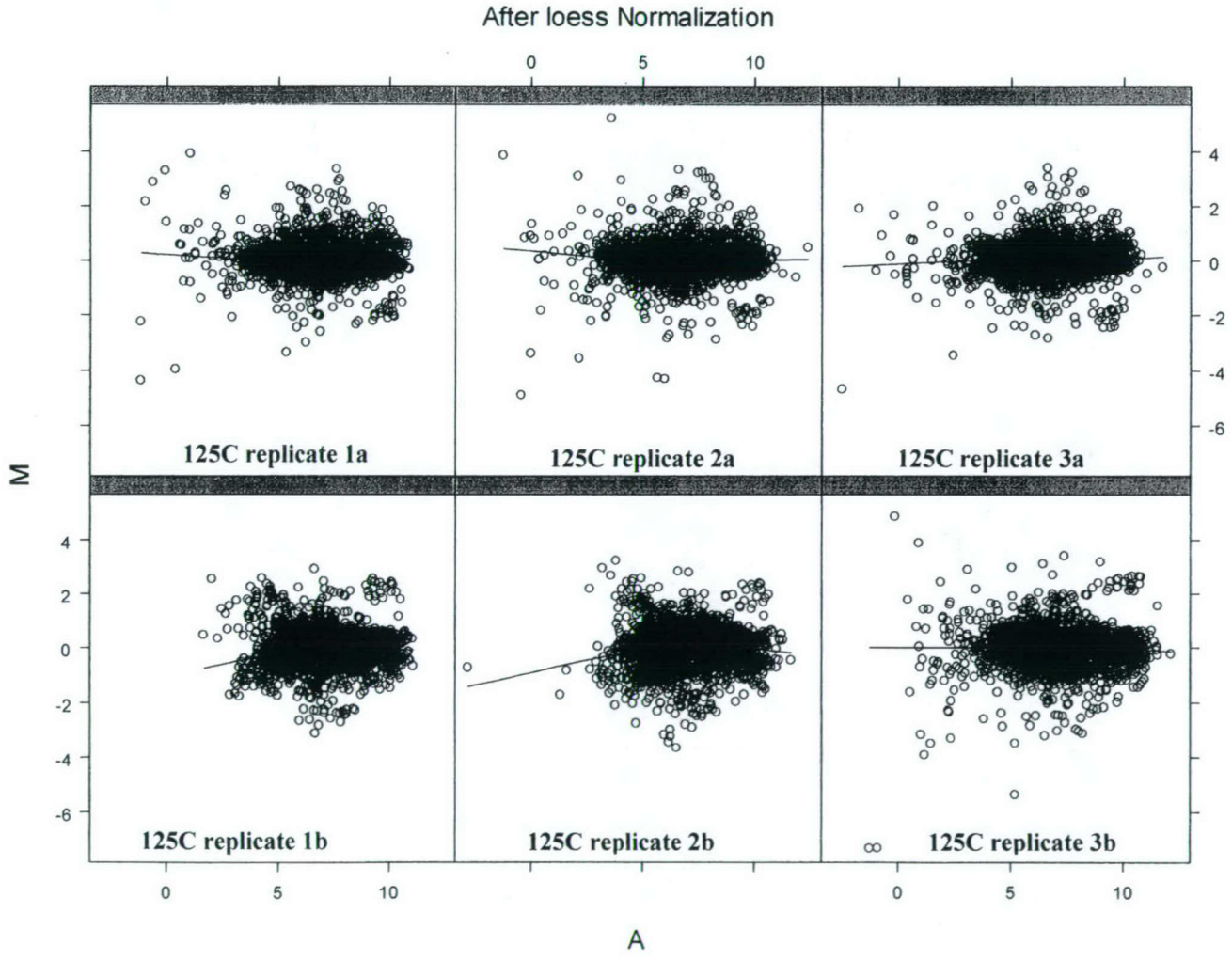


Figure 3-6: Normalization of 125C array data (Replicates represent dye swap experiments):

Box plot illustrate the normalization of data across the replicate arrays. $M = \log_2(Cy5/Cy3)$.

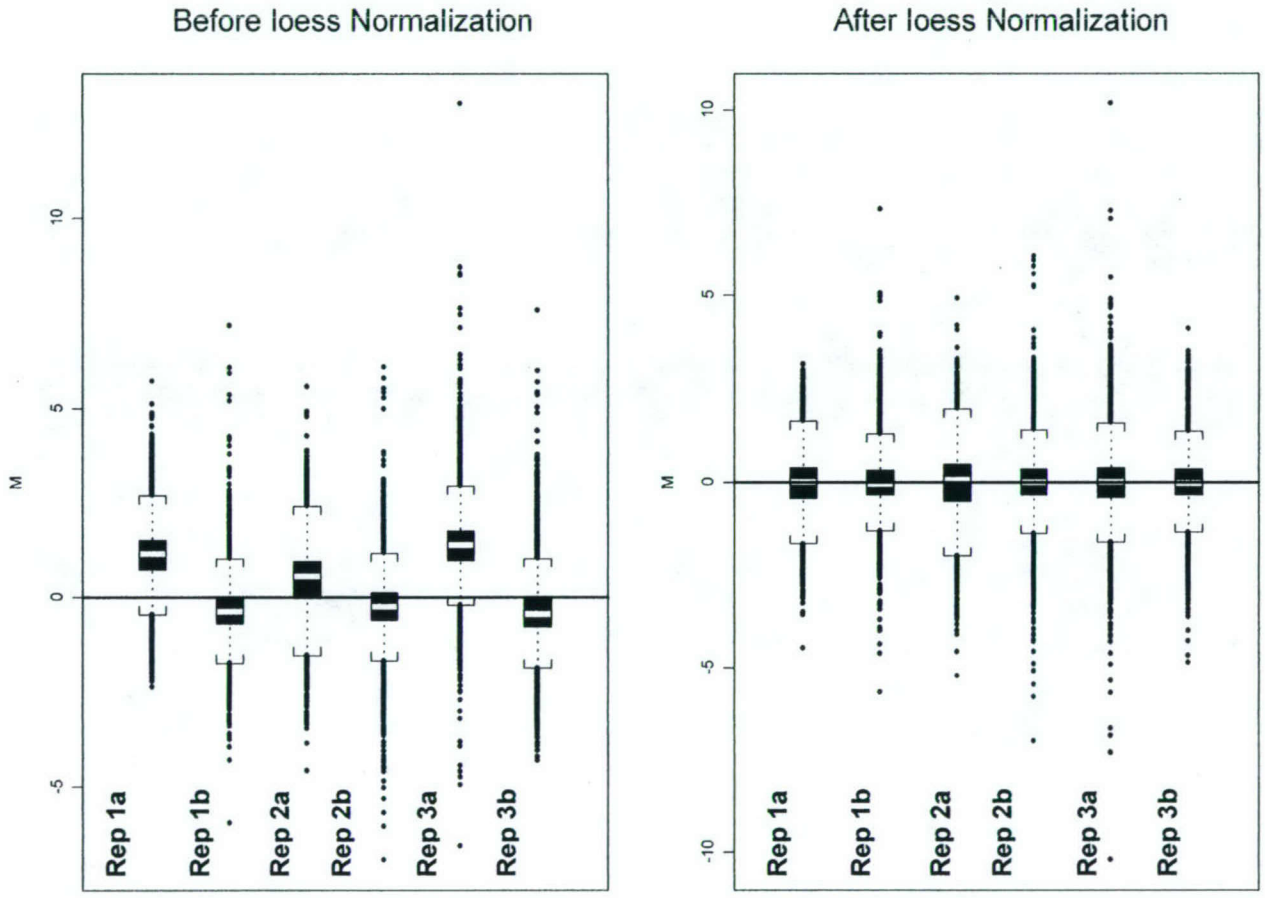


Figure 3-7: Normalization of 125D array data (Replicates represent dye swap experiments):

MvA plot illustrates the normalization of data across the replicate arrays. $M = \log_2(Cy5/Cy3)$. $A = \log_{10}(Cy5*Cy3)$.

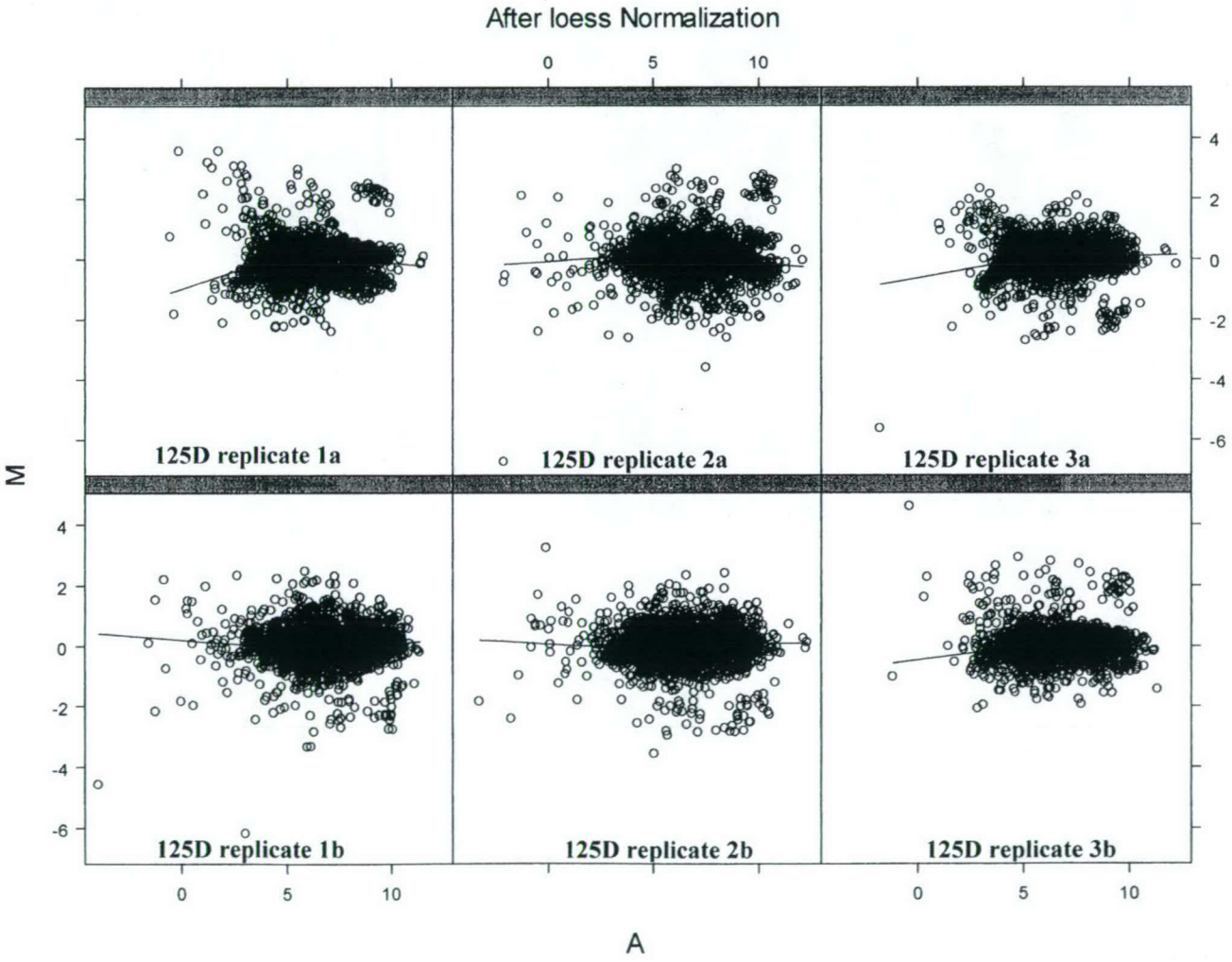


Figure 3-8: Normalization of 125D array data (Replicates represent dye swap experiments):

Box plot illustrate the normalization of data across the replicate arrays. $M = \log_2(Cy5/Cy3)$.

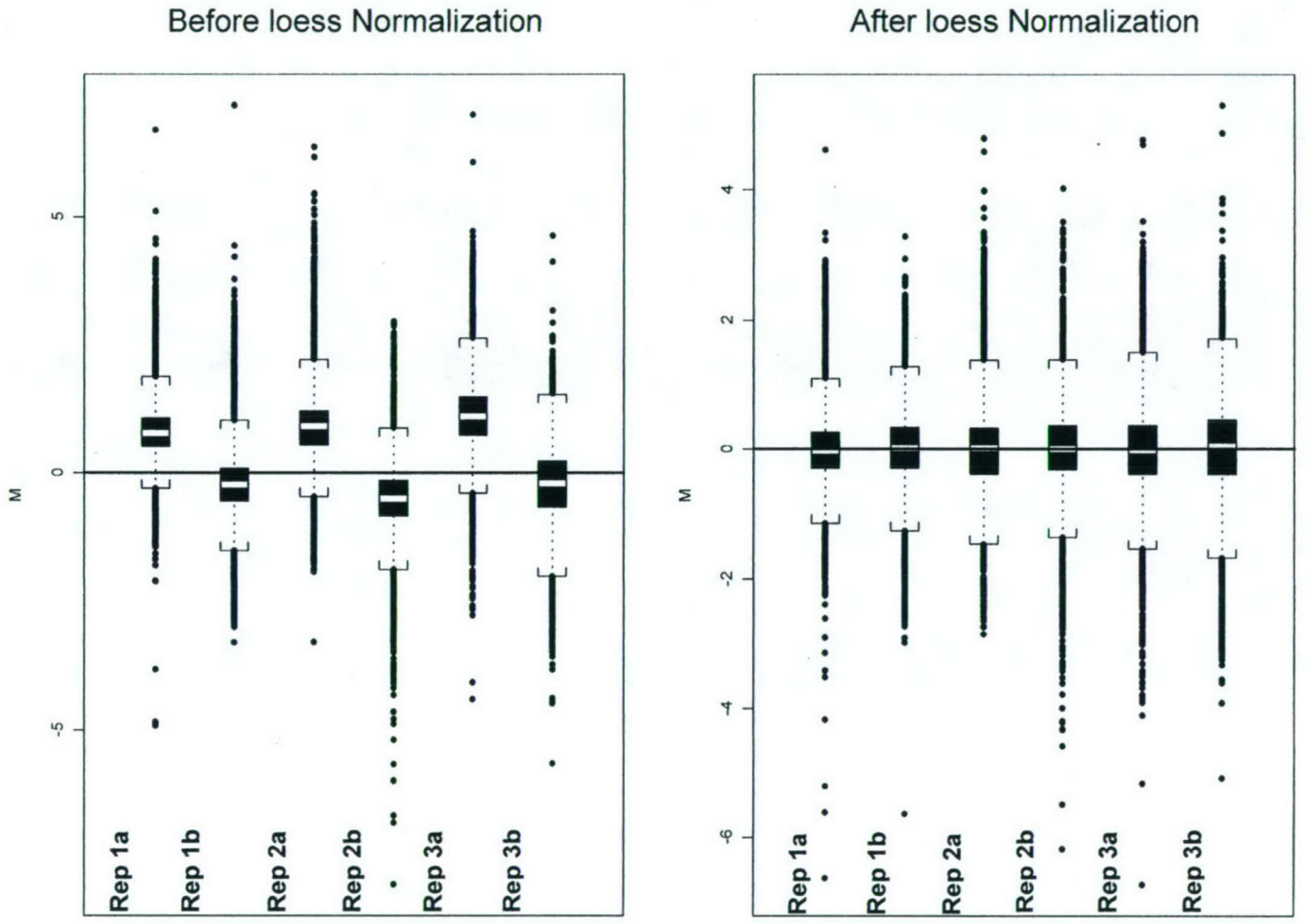


Figure 3-9: Linear regression analysis of control data spots – Experiment AX1.

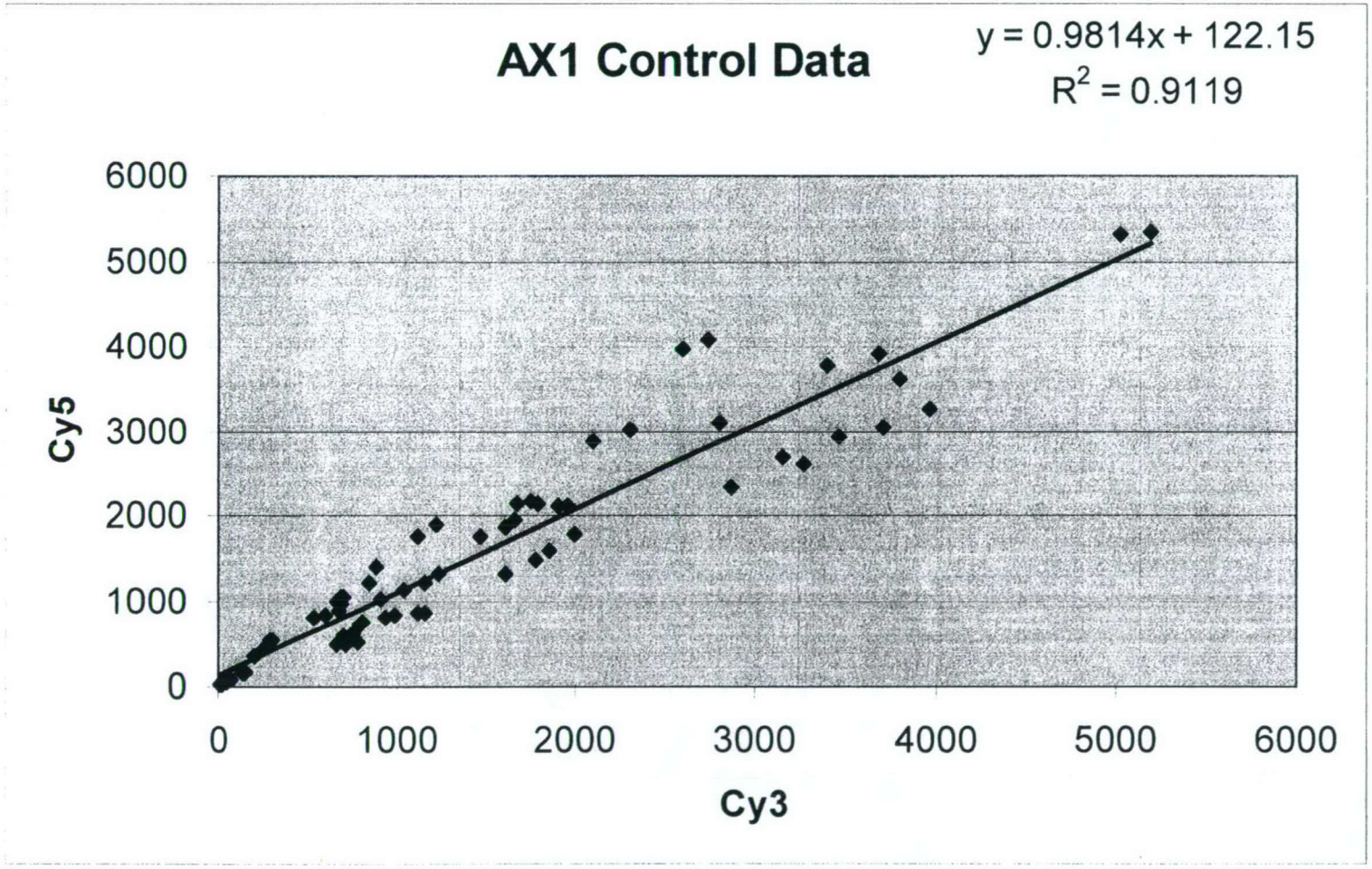


Figure 3-10: Linear regression analysis of control data spots - Experiment 125C.

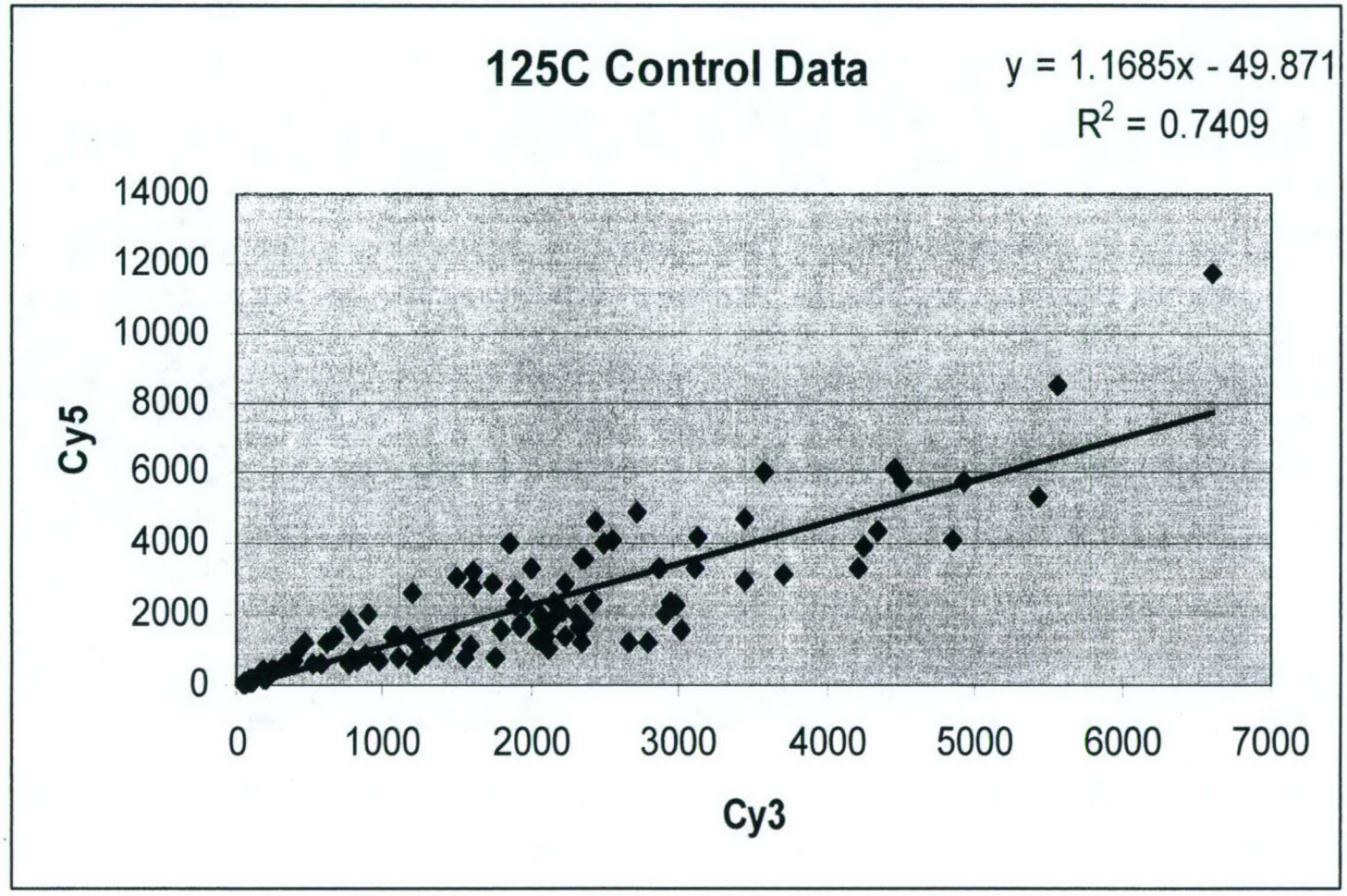


Figure 3-11: Linear regression analysis of control data spots - Experiment 125D.

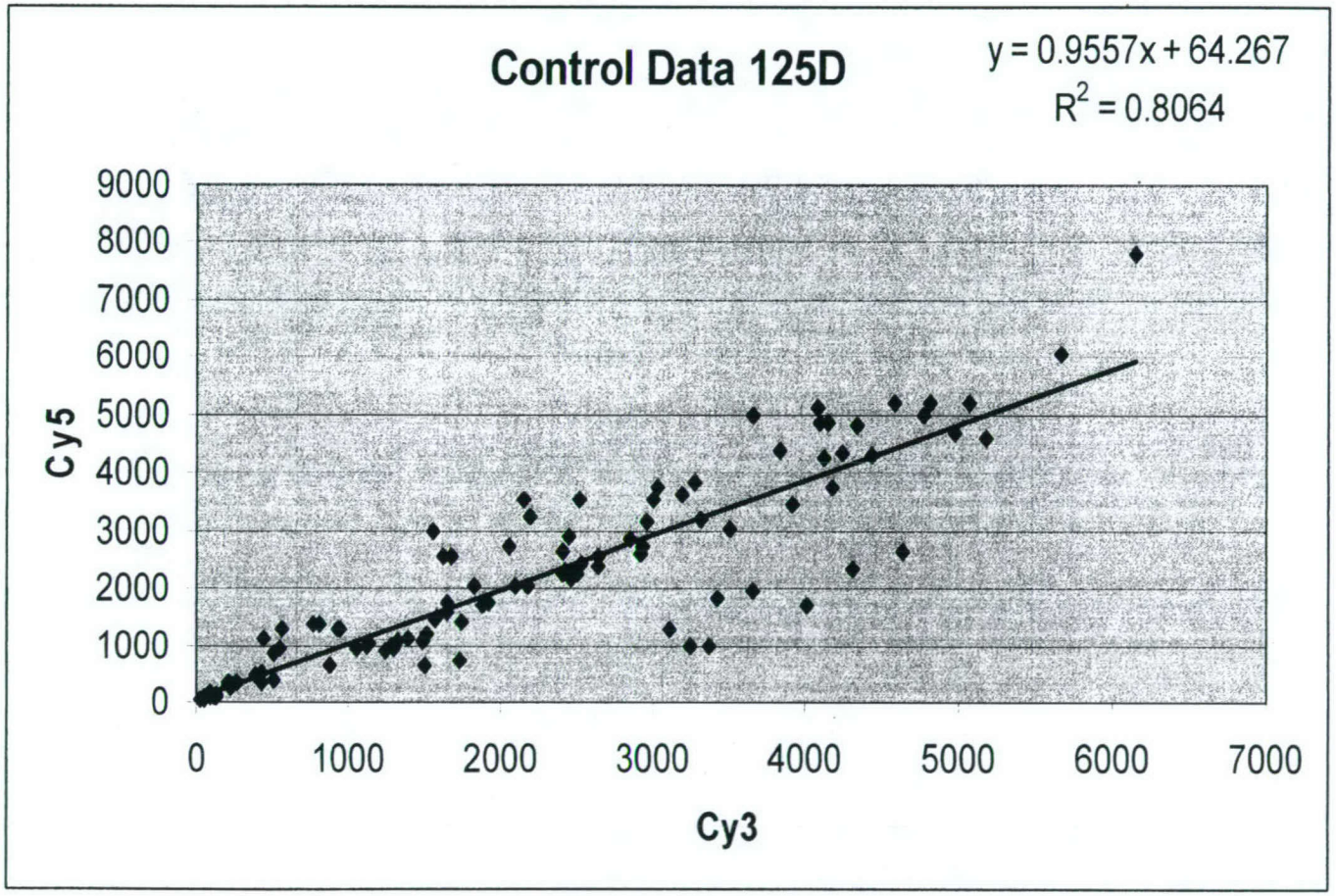


Figure 3-12: AX1 SAM plot (Statistical Analysis of Microarrays) Scatter plots of the observed relative difference d(i) vs. the expected relative difference $d_E(i)$. For the majority of genes, d(i) approximates $d_E(i)$, but some genes are represented by points displaced from the $d(i) = d_E(i)$ line by a distance greater than a designated threshold, delta (represented by the dotted line). Genes that fall outside the cutoff represented by delta are considered significant

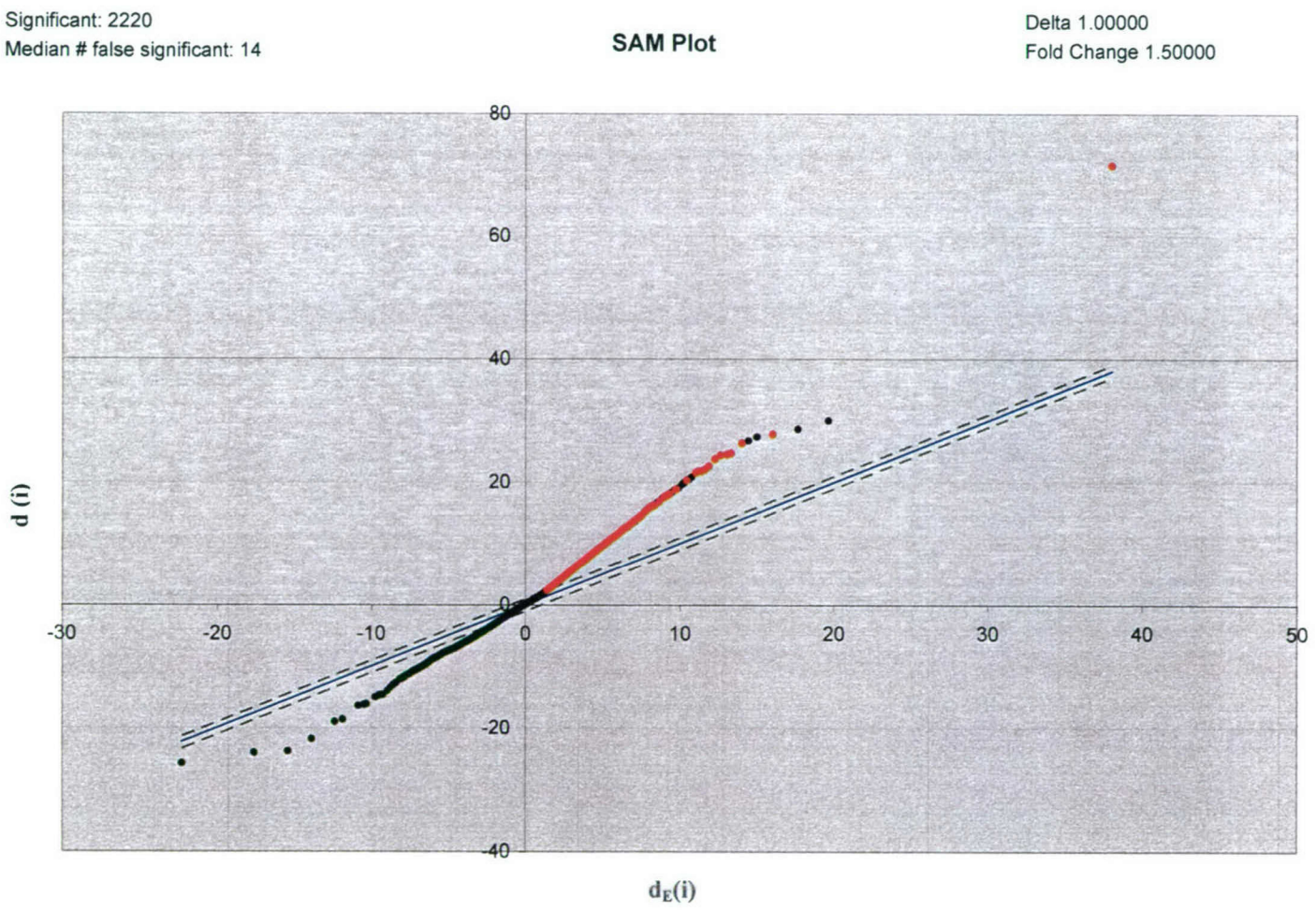


Figure 3-13: 125C SAM plot (Statistical Analysis of Microarrays) Scatter plots of the observed relative difference d(i) vs. the expected relative difference $d_E(i)$. For the majority of genes, d(i) approximates $d_E(i)$, but some genes are represented by points displaced from the $d(i) = d_E(i)$ line by a distance greater than a designated threshold, delta (represented by the dotted line). Genes that fall outside the cutoff represented by delta are considered significant

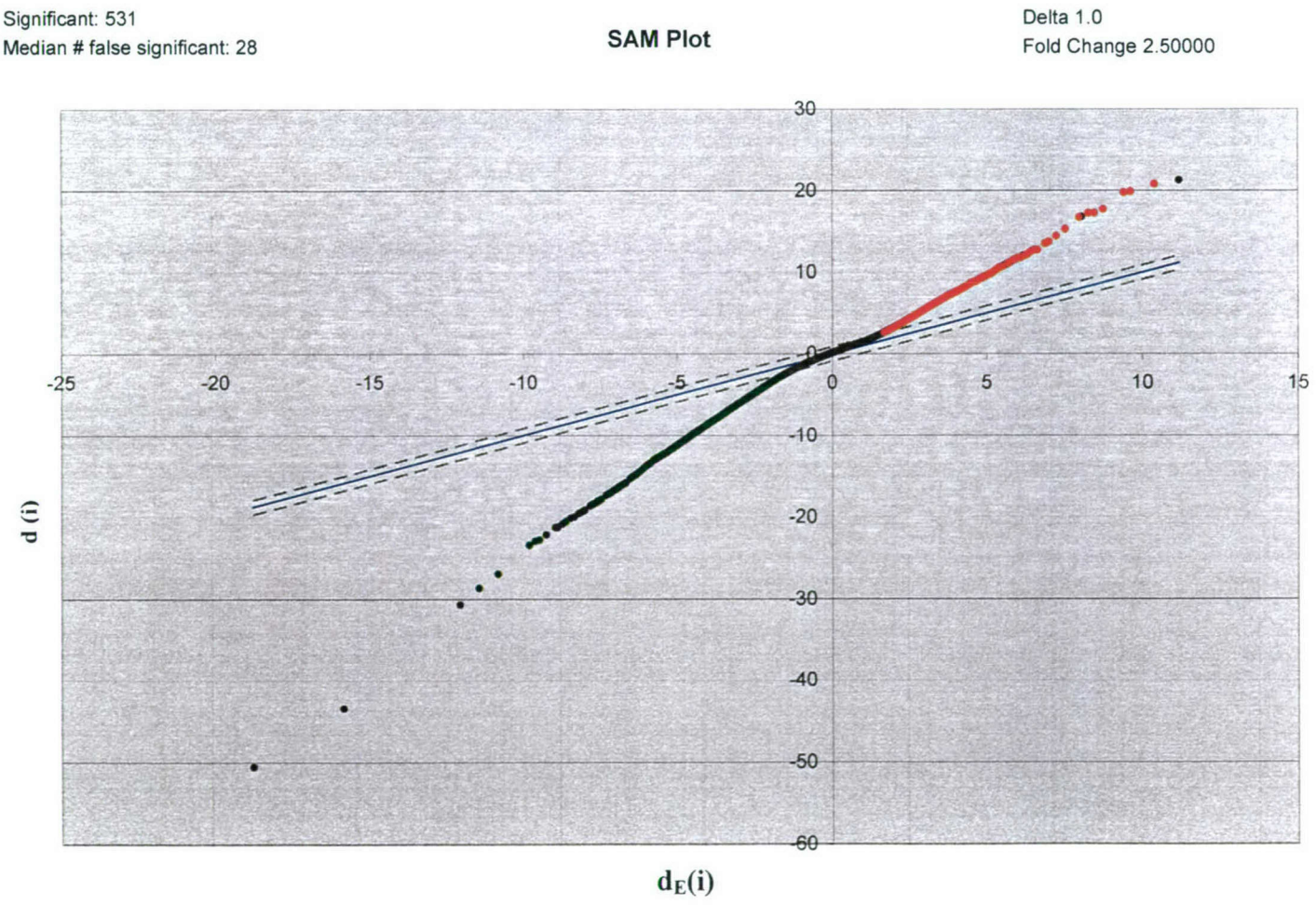
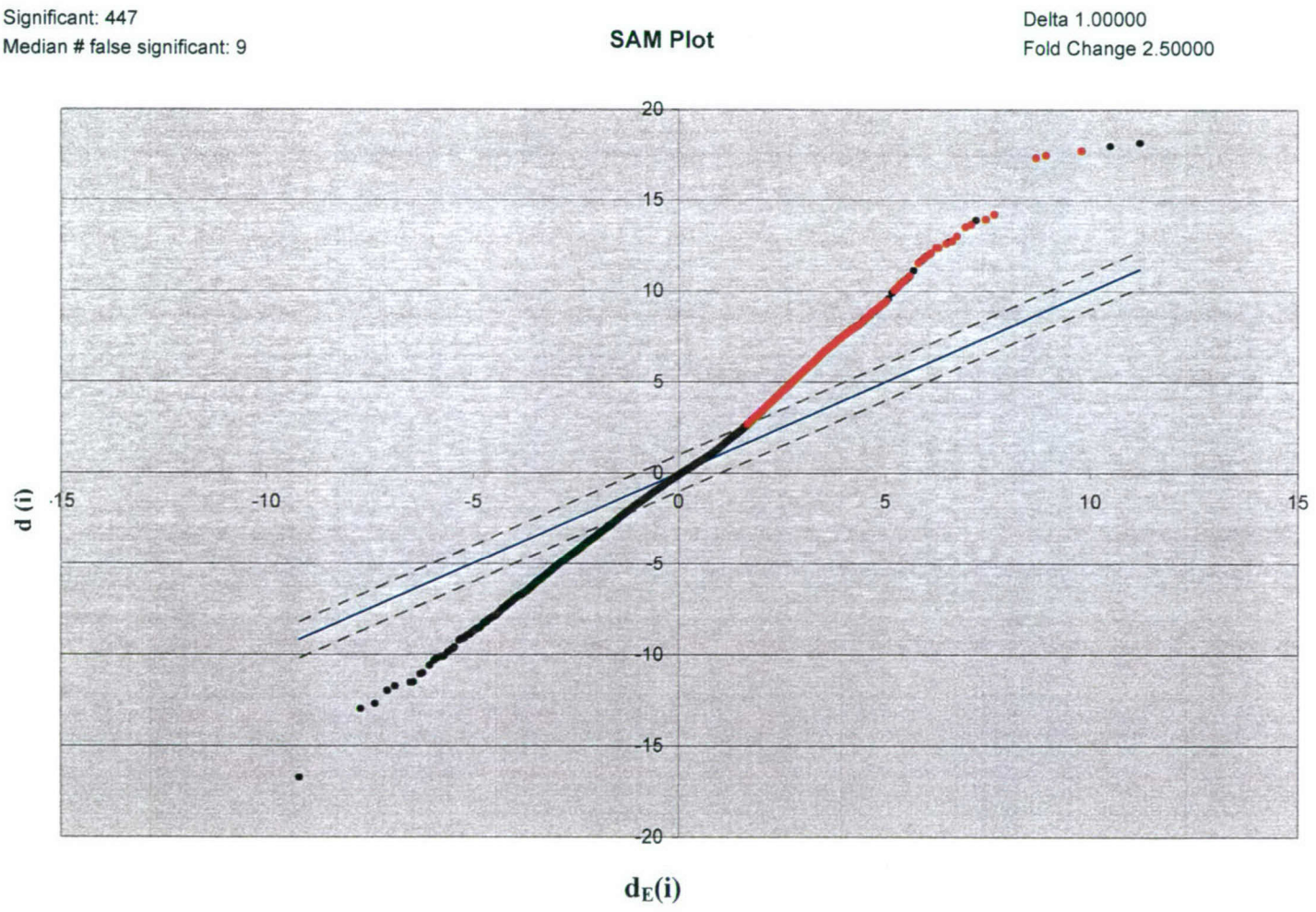


Figure 3-14: 125D SAM plots (Statistical Analysis of Microarrays) Scatter plots of the observed relative difference d(i) vs. the expected relative difference $d_E(i)$. For the majority of genes, d(i) approximates $d_E(i)$, but some genes are represented by points displaced from the $d(i) = d_E(i)$ line by a distance greater than a designated threshold, delta (represented by the dotted line). Genes that fall outside the cutoff represented by delta are considered significant



false discovery rate at which the gene is called significant. The false discovery rate (FDR) is an estimate of the percentage of genes identified by chance that is based on analyzing permutations of the repeated measurements of expression for each gene. By varying delta, the false discovery rate is adjusted. In addition, a second approach to identifying significant gene expression changes is to account for consistent changes between paired samples at a certain fold-change cut-off. In the present study, delta value and fold-change cut-offs were selected to keep the FDR below 2.5%, which means that for every 100 genes called significant, fewer than 2.5 genes would be identified incorrectly.

Initially, each dataset corresponding to the separate growth experiments (125C, 125D, and AX1) was analyzed independently. The fold-change differences in the non-axenic growth experiments were consistently higher than the fold-change differences in the axenic growth experiment. For example, the cDNAs with positive fold-change differences averaged 4.07 ± 0.97 in growth Experiment 125D, 3.85 ± 1.17 in growth Experiment 125C, and 1.92 ± 0.54 in the axenic growth Experiment AX1. Features with a fold-change difference of at least 2.5 or less than 0.5 were considered differentially expressed for Experiments 125C and 125D, while features of at least 1.25 or less than 0.8 were considered differentially expressed for Experiment AX1. By raising the delta value, lower fold-change differences could be tested for statistical significance, which was useful in further analysis below.

Up-regulated and down-regulated cDNAs were compared across the three experiments and only those transcripts that were up-regulated in all three experiment were analyzed further. First, individual cDNAs were identified as up-regulated across all three experiments, next the cDNAs were annotated based on sequence description data previously recorded in analysis of the *P. multiseries* EST database, then the cDNAs were identified as singletons or part of a larger contig. Most of the cDNAs that were differentially expressed were part of a larger contig, therefore, the array data for all of the cDNAs within each contig were analyzed to verify the overall expression of the gene. Encouragingly, the array data confirmed that the individual cDNAs within each contig

were consistently up- or down- regulated. Generally the data fell within the original guidelines for fold-change differences. Occasionally, however, lower fold-change differences were observed that fell within the range of statistical stringency given above (FDR < 2.5%). Any features that were not statistically significant are represented as '**' in the corresponding data table.

Replicate spots may be collapsed by averaging and then running statistical tests, or the replicates may be analyzed as uncollapsed data. In this study, the two layers of replicates (replicate cDNAs on the array and replicate hybridizations) were accounted for by first analyzing the replicates spots uncollapsed using the modified t-test described above, and presenting ratios for each differentially expressed cDNA replicate in tables. Then, the replicate spots were averaged and presented with standard deviations. Finally, gene expression ratios from individual cDNAs within a larger contig were averaged with standard deviations. The data confirm the technical and biological replicability among these experiments.

Identification of mRNAs Regulated in DA Producing Conditions: Nucleotide and deduced amino acid sequences were analyzed using NCBI tools (including Blast and ORF Finder), Pfam, SwissProt, the Vector NTI suite from InforMax (Rockville, MD), and LaserGene from DNAStar (Madison, WI).

Results and Discussion:

The goal of this study was to identify transcripts that were up-regulated in *P. multiseries* cells actively producing domoic acid (DA) compared to *P. multiseries* cells not producing DA. Samples for microarray analysis were obtained from three separate growth experiments; DA production increased and peaked during the stationary growth phase in all three experiments (Figures 3-15 to 3-17). Experiments 125C and 125D were performed under non-axenic culture conditions, whereas AX1 was performed under axenic culture conditions. DA concentrations were 33 times higher in the non-axenic growth experiments than in the axenic growth experiments. Higher DA production in the non-axenic growth experiments was expected, based on results from previous studies (Douglas and Bates, 1992; Douglas et al., 1993; Bates et al., 1995; Bates et al., 2003).

Up-regulated genes: In an effort to select for genes that were correlated specifically with DA production and/or cell growth, significantly expressed genes were compared across the three growth experiments and only those transcripts that were up- regulated in all three growth experiments were considered further. Up-regulation of gene expression was observed for 121 individual cDNAs across all three *P. multiseries* growth experiments. 117 of these 121 cDNAs assembled into 8 unique contigs. The remaining 4 cDNAs (singletons) represented cDNA sequences otherwise not represented in the EST dataset. Up-regulated cDNAs represent 2.25% of the clones printed on the *P. multiseries* chip. The functional identities of the up-regulated transcripts were suggested from sequence similarity to encode a 3-carboxymuconate cyclase, phosphoenolpyruvate carboxykinase (ATP-specific), an amino acid transporter, a small heat shock protein, a long-chain fatty-acid-CoA ligase, an aldo/keto reductase, 5 hypothetical proteins, and one potentially novel protein (Table 3-2). The following discussion will focus individually on each of the unique contigs or sequences.

Figure 3-15: Cell growth and DA production, by *Pseudo-nitzschia multiseries* clone CL-125, axenic cultures (AX1). Cells were harvested for RNA extraction on the days labeled with red arrows (Day 9 and Day 42).

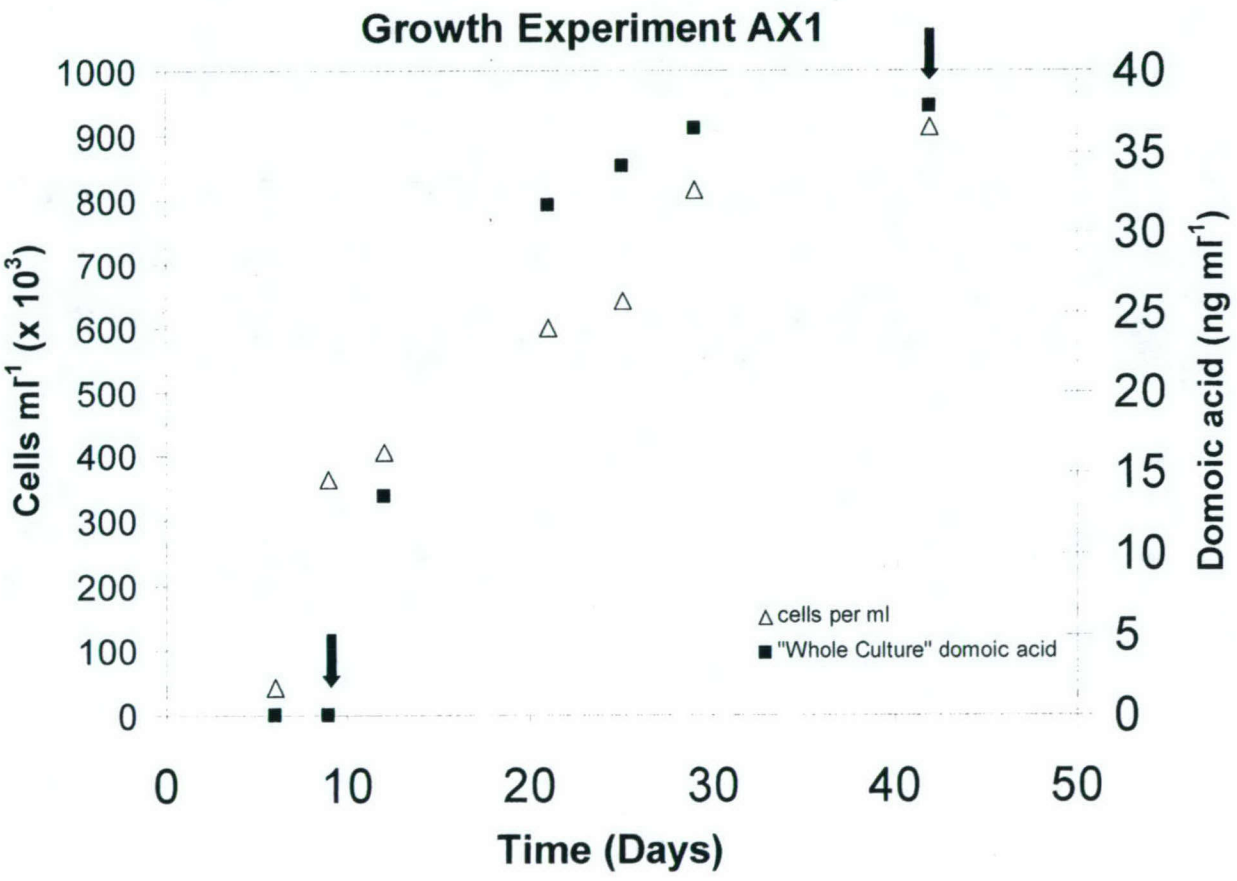


Figure 3-16: Cell growth and DA production, by *Pseudo-nitzschia multiseries* clone CL-125, non-axenic culture (125C). Cells were harvested for RNA extraction on the days labeled with red arrows (Day 4 and Day 10).

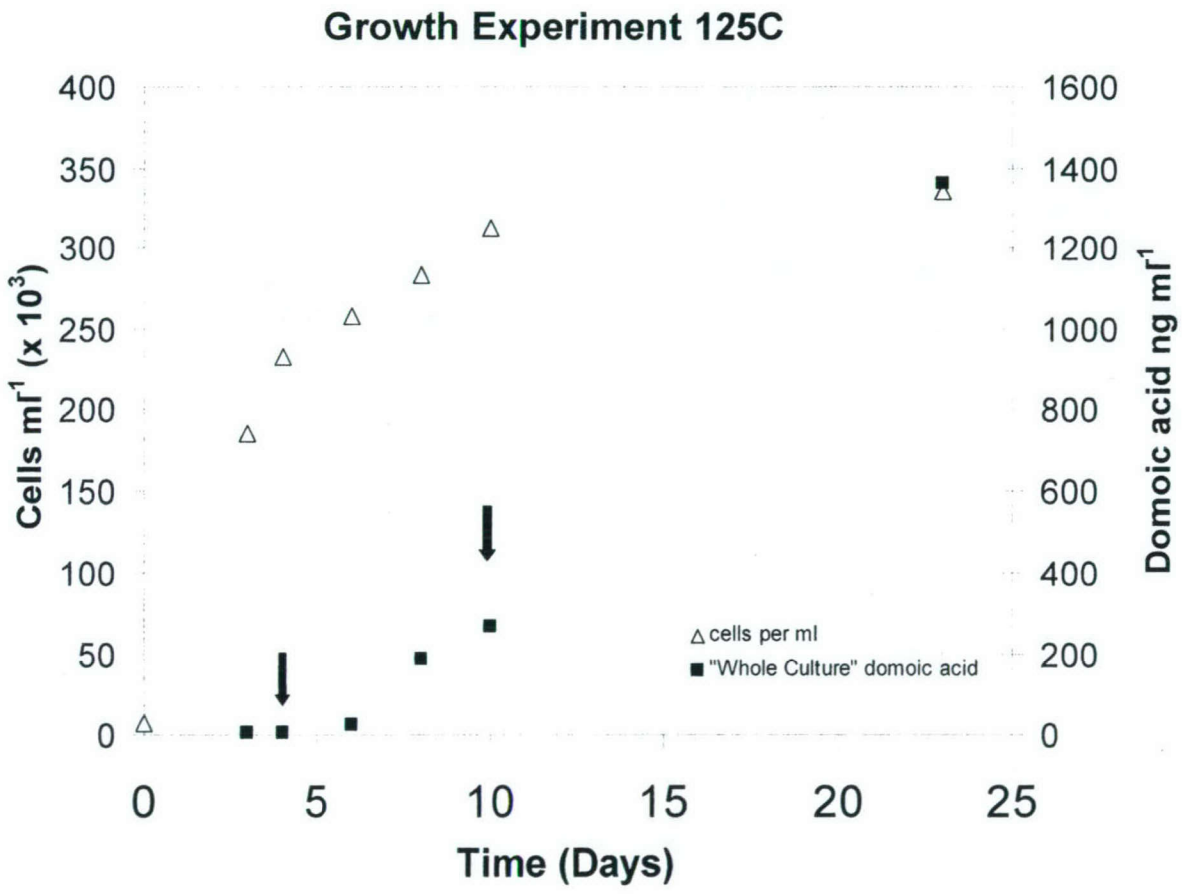


Figure 3-17: Cell growth and DA production, by *Pseudo-nitzschia multiseries* clone CL-125, non-axenic culture (125D). Cells were harvested for RNA extraction on the days labeled with red arrows (Day 4 and Day 10).

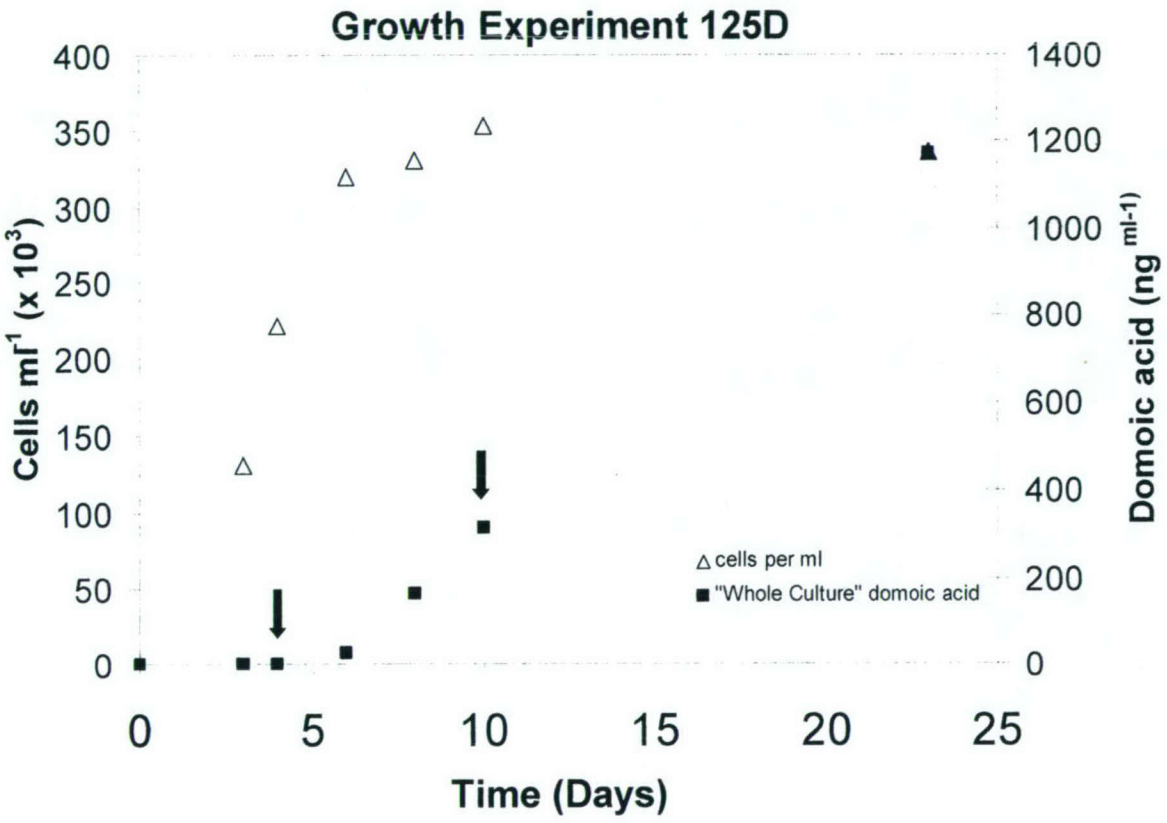


Table 3-2: Overview of Up-regulated cDNAs in PSN Differential Expression Study

PSN Array Identifier	PSN Library Identifier	cDNAs per Contig	(bp)	Putative Identification	Average Fold-change Measurements					
					125D	S.D.	125C	S.D.	AX1	S.D.
Contig 1	PSN0011	22	2236	3-carboxymuconate cyclase	**4.03**	0.80	**3.24**	0.61	**2.16**	0.44
Contig 3	PSN0016	14	2158	Phosphoenolpyruvate carboxykinase (ATP)	**3.78**	0.26	**2.94**	0.42	**2.95**	0.48
Contig 4	PSN0072	3	1818	Na+ and Cl- dependent amino acid transporter	**3.31**	0.70	**3.24**	0.25	**1.91**	0.23
Contig 5	PSN0025	7	1001	Small heat shock protein /Hsp 20	**7.00**	0.16	**7.81**	0.43	**4.10**	0.41
Contig 6	PSN0014	13	2438	Long-chain fatty-acid-CoA ligase	**4.66**	0.92	**3.76**	0.79	**2.13**	0.26
Contig 7	PSN0015	7	1742	Aldo/keto reductase	**3.12**	0.60	**2.88**	0.26	**1.83**	0.18
Contig 8	PSN0042	4	1055	Tyrosine sulfation activity	**5.10**	1.49	**6.90**	0.93	**3.01**	0.87
Contig 2	PSN0002	54	2445	NO HITS	**4.25**	0.71	**4.24**	0.56	**1.57**	0.21
Singleton 1	17F11	1	919	Unknown with possible Hydrolase activity	**5.45**	0.14	**4.98**	1.71	**2.07**	0.08
Singleton 2	75E8	1	552	Unknown with possible 3-O-acyltransferase activity	**3.26**	0.14	**3.21**	0.05	**3.54**	0.07
Singleton 3	6H1	1	1270	Unknown, with possible glutamine-hydrolyzing activity	**5.45**	0.40	**2.81**	0.06	**2.55**	0.06
Singleton 4	45H6	1	323	Unknown, with possible isomerase activity	**3.35**	0.05	**3.88**	0.13	**3.17**	0.05

Contig 1, Cycloisomerase: Contig 1includes 22 cDNAs, which form a consensus sequence 2236 bp long (Figure 3-18). Overall average expression ratios (fold-change differences) were 3.24 (± 0.61) in Experiment 125C, 4.03 (± 0.80) in Experiment 125D, and 2.16 (± 0.44) in Experiment AX1 (Table 3-3). The predicted coding region for Contig 1 revealed an open reading frame (ORF) of 525 amino acids (Figure 3-19), which aligned with COG2706, a cluster of orthologues that identifies a conserved domain for 3-carboxymuconate cyclase (Figure 3-20). Additional BLAST analysis supported the temporary assignment of Contig 1 as a muconate cycloisomerase (Tables 3-4 and 3-5). The specific enzyme that this contig aligns most closely with, carboxy-cis,cis-muconate cyclase, catalyzes the cycloisomerization of 3-carboxy-2,5-dihydro-5-oxofuran-2-acetate to 3-carboxy-cis,cis-muconate (Figure 3-21). This isomerization is reminiscent of that suggested in DA synthesis (Ramsey et al., 1998) and offers a target molecule to focus on which may be directly involved in cyclization leading to the pyrrolidine ring in the DA molecule. Alternatively, the enzyme may be involved in converting aromatic compounds into citric acid cycle intermediates, which have been proposed to feed the pathway leading to DA synthesis (Ramsey et al., 1998). Searching against the *Thalassiosira pseudonana* genome revealed a similar sequence within *T. pseudonana* scaffold 79, which shared 70% identity, and 78% similarity to *P. multiseries* Contig 1 (Figure 3-22).

Figure 3-18:
Sequence Alignment Overview for Contig 1 (2236bp, 22 clones, 35 sequences):
(In the sequence alignment diagrams throughout this document, dotted lines vs. solid lines represent the direction that the cDNAs were sequenced.)

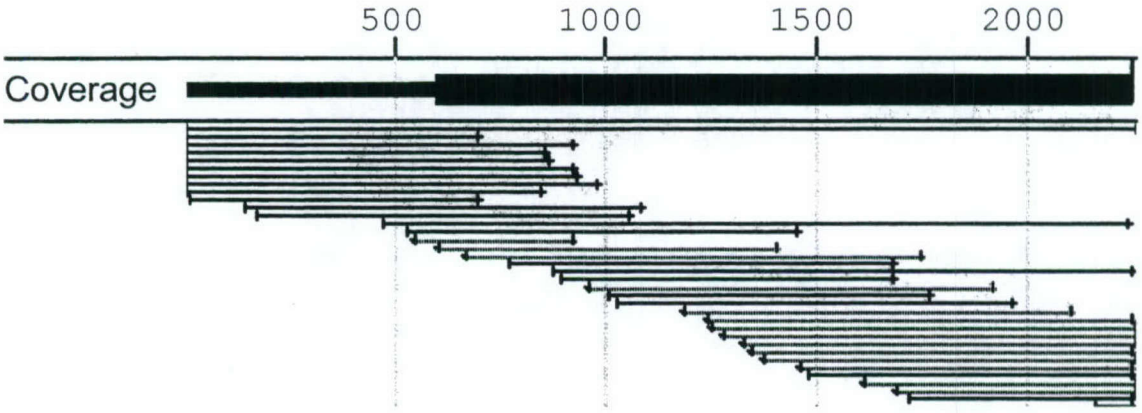


Figure 3-19: Predicted coding region for Contig 1:

Length: 525 aa

Frame	from	to	Length
+2	155..	1732	1578

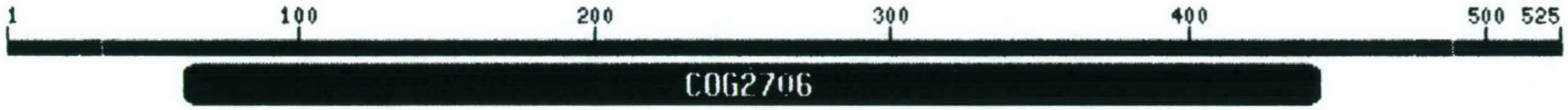
MRIYQRTPTDLSATTAGTFIRSDSNEDEGEDDDHQLFFVTSYSDFEKLAHGPRGHEAKHSVHVYRF
FPSDGSLVLLNIQGDADVVTNPAFSRHHPRLNVIYTCTEDCHENGRIIAFKVKPDGTLEQFGEPVDA
GGTSTCYLTIDKAERNLLAVNYWNSTLVVIPMDPDTGALIGGVKNVYDPNMGKTMVACAKKDG
GVNHSCNDASTISARQADPHSHALVLDPFVGRVAYVPDLGKDLVREFYYDATEGNIAIELNVMPS
GLCTGQPDGPRYLDFHPEYNIAYVVNELSSTVAVFEVDRELLNEIHEASRNGEDMNRFRGRSTLRL
VQSIKTIPHAFPTTMNTCGRMCVHKSGRYVIVSNRGHQSITVFRVKTKGSKRGELQIVGCYHTRGE
TPRHFQFDNSGQYLLVANQDTDSIAVFNFNLSNGELKYSGNEYRVPSPNFVCCCPTYSEDDTEIRQ
RQENFESSIRAVTLAKDNENNSGSDSEDSTVPTWRGRSSEDNIKAELAKAREEIETLKKLLAERVQ

Table 3-3: Fold-change Measurements for Individual cDNAs within Contig 1:

	cDNA Identifier	125D replicate spot 1	125D replicate spot 2	125D Average	S.D.	125C replicate spot 1	125C replicate spot 2	125C Average	S.D.	AX1 replicate spot 1	AX1 replicate spot 2	AX1 Average	S.D.
1	135E8	2.83	3.22	3.03	0.28	6.39	3.45	4.92	2.08	2.59	2.46	2.53	0.09
2	136D7	4.44	4.20	4.32	0.17	3.27	3.23	3.25	0.03	2.28	2.24	2.26	0.03
3	163D4	4.47	4.70	4.58	0.17	3.83	3.78	3.81	0.04	2.44	2.38	2.41	0.04
4	169H10	5.25	4.97	5.11	0.19	3.73	3.73	3.73	0.00	2.77	2.77	2.77	0.00
5	180E5	5.24	4.49	4.86	0.53	3.27	3.30	3.28	0.02	2.43	2.46	2.45	0.02
6	183G5	2.23	2.19	2.21	0.03	**	**	**	**	1.48	1.46	1.47	0.02
7	45G9	4.23	4.23	4.23	0.00	3.08	2.60	2.84	0.34	2.33	2.23	2.28	0.07
8	51A4	3.51	3.32	3.41	0.14	3.41	3.36	3.38	0.03	2.75	**	2.75	----
9	51F10	4.38	4.45	4.41	0.04	3.22	3.30	3.26	0.06	2.11	2.07	2.09	0.02
10	51F9	4.91	4.94	4.93	0.02	3.33	3.45	3.39	0.09	2.52	2.50	2.51	0.01
11	52B1	4.10	3.99	4.05	0.08	2.47	2.24	2.35	0.16	1.63	1.67	1.65	0.03
12	52E12	4.81	**	4.81	-----	3.73	3.62	3.68	0.08	2.21	2.33	2.27	0.08
13a	54B7 rep1	3.46	3.54	3.50	0.06	2.64	2.67	2.65	0.02	1.80	1.76	1.78	0.03
13b	54B7 rep2	2.95	2.78	2.86	0.12	2.85	2.72	2.78	0.09	1.76	1.80	1.78	0.02
14	55C4	3.48	3.59	3.53	0.08	3.63	3.35	3.49	0.19	2.17	2.17	2.17	0.00
15	57F7	4.08	3.97	4.03	0.08	2.95	2.96	2.95	0.00	2.41	2.46	2.44	0.03
16	6F2	4.48	4.51	4.50	0.02	3.28	3.34	3.31	0.05	1.30	1.33	1.31	0.02
17	71H1	4.91	5.10	5.00	0.14	3.50	3.51	3.50	0.00	2.74	2.68	2.71	0.05
18	71H3	2.99	4.03	3.51	0.73	2.72	2.63	2.68	0.06	1.94	2.07	2.00	0.09
19	76D7	3.55	3.71	3.63	0.11	2.23	2.28	2.26	0.04	1.53	1.57	1.55	0.03
20	180D8	**	**	**		**	**	**	**	1.90	1.88	1.89	0.02
	AVERAGE	4.02	4.00	**4.03**		3.34	3.13	**3.24**		2.15	2.11	**2.16**	
	S.D.	0.85	0.77	**0.80**		0.86	0.47	**0.61**		0.44	0.41	**0.44**	

** = Not significant

Figure 3-20: Contig 1 Sequence Alignment with COG2706, Conserved Domain for 3-carboxymuconate cyclase:



		Score (bits)	E value
gnl|CDD|12130,	COG2706, COG2706, 3-carboxymuconate cyclase [Carbohydrate transport and metabolism]	185	8e-48

Conserved Domain -Length = 346 residues, 93.6% aligned
Score = 185 bits (472), Expect = 8e-48

```
Query:  61   VHVYRFFPSDGSLVLLNIQGDADVVTNPAFSRHHPRLNVIYTCTEDCHENGRIIAFKVKP  120
Sbjct:  18   IYVFNLDTKTGELSLLQL---VAELGNPTYLAVNPDQRHLYVVNEPGEE-GGVAAYRIDP  73

Query:  121  -DGTLEQFGEPVDAGGTSTCYLTIDKAERNLLAVNYWNSTLVVIPMDPDTGALIGGVKNV  179
Sbjct:  74   DDGRLTFLNR-QTLPGSPPCYVSVDEDGRFVFVANYHSGSVSVYPLQAD-GSLQPVVQVV  131

Query:  180  YDPNMGKTMVACAKKDGGVNHSCNDASTISARQADPHSHALVLDPFVGRVAYVPDLGKDL  239
Sbjct:  132  KH----------------------TGSGPHERQESPHVHSANFTP-DGRYLVVPDLGTDR  168

Query:  240  VREFYYDATEGNIAIELNVMPSGLCTGQPDGPRYLDFHPEYNIAYVVNELSSTVAVFEVD  299
Sbjct:  169  IFLYDLDDGKLTPADPAEVKPGA-------GPRHIVFHPNGKYAYLVNELNSTVDVLEYN  221

Query:  300  RELLNEIHEASRNGEDMNRFRGRSTLRLVQSIKTIPHAFPTTMNTCGRMCVHKSGRYVIV  359
Sbjct:  222  PAV--------------------GKFEELQTIDTLPEDFTGT-NWAAAIHISPDGRFLYA  260

Query:  360  SNRGHQSITVFRVKTKGSKRGELQIVGCYHTRGETPRHFQFDNSGQYLLVANQDTDSIAV  419
Sbjct:  261  SNRGHDSIAVFSVDPDG---GKLELVGITPTEGQFPRDFNINPSGRFLIAANQKSDNITV  317

Query:  420  FNFNLSNGELKYSGNEYRVPSPNFVC  445
Sbjct:  318  FERDKETGRLTLLGRYAVVPEP--VC  341
```

Table 3-4: Contig 1 Blast Results, Overview:

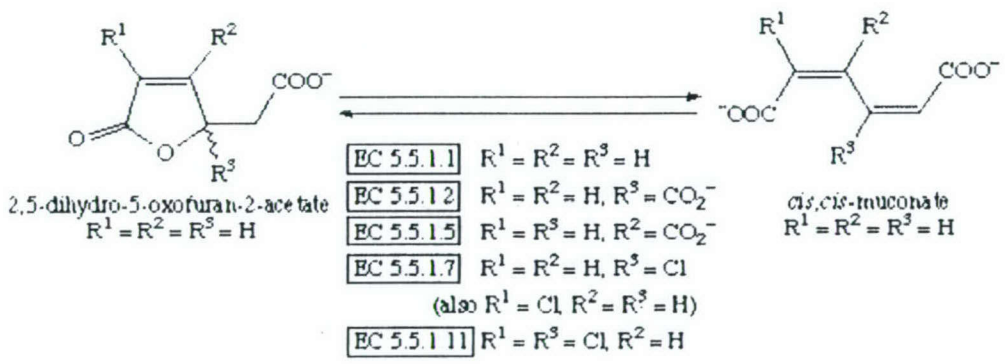
PSN Identifier	NCBI Database	Consensus length (bp)	NCBI Identifier	Putative Identification	*Species* or Domain Name	E-value
Contig 1	CDD	2236	NP_865629.1	3-carboxymuconate cyclase	COG2706	8.00E-48
Contig 1	NR	2236	NP_865629.1	cycloisomerase	Pirellula sp.	1.00E-27

Table 3-5: Contig 1, Individual cDNA BlastX Results - NR, Overview:

	PSN Identifier	Length (bp)	NCBI Identifier	Putative Identification	*Species* or Domain Name	E-value
1	135E8	(M13r - 948, T7 - 978)	ZP_00064353.1	3-carboxymuconate cyclase	Leuconostoc mesenteroides	9.00E-06
2	136D7	901		NO HITS		
3	163D4	1293	ZP_00064353.1	3-carboxymuconate cyclase	Leuconostoc mesenteroides	2.00E-13
4	169H10	(M13r - 832, T7 - 952)	NP_865629.1	putative cycloisomerase	Pirellula sp.	2.00E-04
5	180E5	(M13r - 871, T7 - 930)	NP_865629.1	putative cycloisomerase	Pirellula sp.	2.00E-06
6	183G5	364				> e-4
7	45G9	1671	ZP_00064353.1	3-carboxymuconate cyclase	Leuconostoc mesenteroides	1.00E-13
8	51A4	(M13r - 639, T7 - 879)				> e-4
9	51F10	711	NP_865629.1	putative cycloisomerase	Pirellula sp.	1.00E-16
10	51F9	918	NP_865629.1	putative cycloisomerase	Pirellula sp.	2.00E-06
11	52B1	736	NP_865629.1	putative cycloisomerase	Pirellula sp.	7.00E-20
12	52E12	(M13r - 839, T7 - 861)	ZP_00068414.1	3-carboxymuconate cyclase	Microbulbifer degradans	5.00E-04
13	54B7	769	NP_865629.1	putative cycloisomerase	Pirellula sp.	5.00E-15
14	55C4	738				> e-4
15	57F7	552		NO HITS		
16	6F2	(M13r - 936, T7 - 897)	NP_865629.1	putative cycloisomerase	Pirellula sp.	2.00E-20
17	71H1	597				> e-4
18	71H3	886				> e-4
19	76D7	878	NP_865629.1	putative cycloisomerase	Pirellula sp.	4.00E-14
20	180D8	941				> e-4
21	164E9	844	NP_865629.1	putative cycloisomerase	Pirellula sp.	5.00E-12
22	24E5	670				> e-4

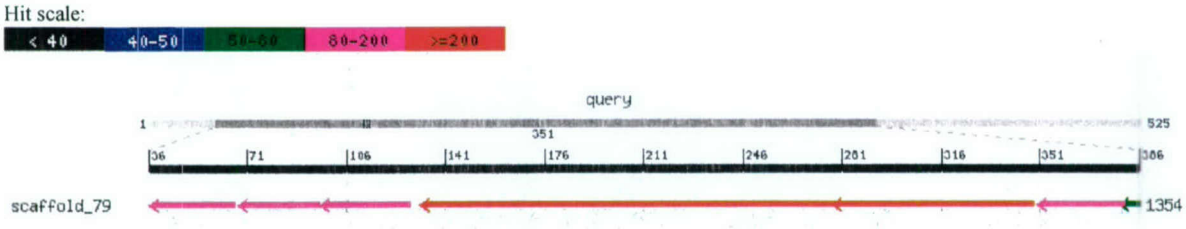
Figure 3-21: 3-carboxymuconate cyclase - Reaction catalyzed:

3-carboxy-2,5-dihydro-5-oxofuran-2-acetate <=> 3-carboxy-cis,cis-muconate



Reference Reaction obtained from: **KEGG: Kyoto Encyclopedia of Genes and Genomes http://www.genome.ad.jp/.

Figure 3-22: Contig 1 Sequence Alignment with *Thalassiosira Pseudonona*:




```
Thal :    40230 FDTDGSLVLLHVAGEGALVINPAFSRFHPR 40141
                F +DGSLVLL++ G+  +V NPAFSR HPR
Contig 1:    67 FPSDGSLVLLNIQGDADVVTNPAFSRHHPR 96

Thal :    39946 RLNVVYTCTEDIEENGQILAYEINTDGSLVEIG 39848
                RLNV+YTCTED  ENG+I+A+++  DG+L + G
Contig 1:    96 RLNVIYTCTEDCHENGRIIAFKVKPDGTLEQFG 128

Thal :    39844 VDAGGTSTCYITIDKDQKHLIAVNYWDSSLVVIPLSKETGGFAGPIQNTY 39695
                VDAGGTSTCY+TIDK +++L+AVNYW+S+LVVIP+  +TG   G ++N Y
Contig 1:   131VDAGGTSTCYLTIDKAERNLLAVNYWNSTLVVIPMDPDTGALIGGVKNVY 180

Thal :    39694 DPRQGLAMKAAAKKHGGANHSNNDDTTIAQRQVDPHSHALVLDPFEGCVA 39545
                DP  G  M A AKK GG NHS ND +TI+ RQ DPHSHALVLDPF G VA
Contig 1:   181 DPNMGKTMVACAKKDGGVNHSCNDASTISARQADPHSHALVLDPFVGRVA 230

Thal :    39544 YVPDLGKDLVREFWYDKEGGKIGSELNVLPSGLSTGKADGPRYFEFHPRF 39395
                YVPDLGKDLVREF+YD   G I  ELNV+PSGL TG+ DGPRY +FHP +
Contig 1:   231 YVPDLGKDLVREFYYDATEGNIAIELNVMPSGLCTGQPDGPRYLDFHPEY 280

Thal :    39394 NVAYVVNELSSTVS 39353
                N+AYVVNELSSTV+
Contig 1:   281 NIAYVVNELSSTVA 294


Thal :    39325 HSLHSLVANITVLLQ*VAVFRVDRELLSEIARAAKLKQPMDKFKGRSTLK 39176
                H  +++   +  L   VAVF VDRELL+EI  A++  + M++F+GRSTL+
Contig 1:   277 HPEYNIAYVVNELSSTVAVFEVDRELLNEIHEASRNGEDMNRFRGRSTLR 326

Thal :    39175 LIQSIKTVPSAFPTQMNTCGRIC 39107
                L+QSIKT+P AFPT MNTCGR+C
Contig 1:   327 LVQSIKTIPHAFPTTMNTCGRMC 349

Thal :    39000 VHQSGRFVIVSNRGHESITIFRVKQSNGSSGK 38905
                VH+SGR+VIVSNRGH+SIT+FRVK      G+
Contig 1:   350 VHKSGRYVIVSNRGHQSITVFRVKTKGSKRGE 381

Thal :    38898 KGSLAQVGFFHTRGET 38851
                +G L  VG +HTRGET
Contig 1:   379 RGELQIVGCYHTRGET 394


Thal :    38770 PRHFQFDHSGQFLIVANQDSDTIAVFSFNLTSGEIKYTGNEYRVPSPNFV 38621
                PRHFQFD+SGQ+L+VANQD+D+IAVF+FNL++GE+KY+GNEYRVPSPNFV
Contig 1:    95 PRHFQFDNSGQYLLVANQDTDSIAVFNFNLSNGELKYSGNEYRVPSPNFV
```

Contig 3, Phosphoenolpyruvate carboxykinase: Contig 3 includes 14 cDNAs, which align to form a consensus sequence 2158 bp long (Figure 3-23). Overall average expression ratios were 3.78 (± 0.26) in Experiment 125C, 2.94 (± 0.42) in Experiment 125D, and 2.95 (± 0.48) in Experiment AX1 (Table 3-6). The predicted coding region for Contig 3 revealed an open reading frame (ORF) of 532 amino acids (Figure 3-24). Blast analysis of the Contig 3 ORF against the SwissProt database revealed a highly significant hit against pfam01293 (E-value = 4e-173) and COG1866 (E-value = 0.0), both clusters of orthologues that code for phosphoenolpyruvate carboxykinase (PCK) (Figure 3-25). The consensus nucleotide sequence and individual cDNAs within the contig were also blasted against the NR database and results supported the putative assignment of Contig 3 as PCK (Tables 3-7 and 3-8). Alignment of the deduced protein with known PCKs revealed 70% similarity and 58% identity with PCK of *Campylobacter jejuni*, 66% similarity, 55% identity with COG1866, 67% similarity, 51% identity with PCK of *Escherichia coli*, 62% similarity, 46% identity with PCK of *Saccharomyces cerevisiae*, and 62 % similarity, 45% identity with PCK of *Arabidopsis thaliana*. Searching the *T. pseudonana* genome database revealed a similar sequence with 77% similarity, 73% identity (Figure 3-26).

Two isoforms of PCK exist, which catalyze either ATP-dependent or GTP-dependent decarboxylation of oxaloacetate into phosphoenolpyruvate (PEP) (Figure 3-27). Contig 3 aligned with ATP-dependent PCK, which may be involved in several functions, including gluconeogenesis, pyruvate metabolism, and C4 photosynthesis (Figures 3-28, -29, -30). PCK may play a role in anaplerotic formation of 2-oxoglutarate, leading to the synthesis of a glutamate derivative (Lea et al., 2001). The glutamate derivative could then lead to domoic acid synthesis as suggested in both of the DA models (Ramset et al., 1998; Smith et al., 2001). Alternatively, the supply of pyruvate could contribute to isoprenoid metabolism. Ramsey et al. (1998) suggest that the principal pathway to the isoprenoid portion of DA is via an alternative route from the traditional acetate-mevalonate pathway to isoprenoid synthesis, which utilizes glyceraldehyde 3-phosphate (GAP) and pyruvate (Eisenreich, et al., 1998). The supply

carbon dioxide via PCK expression may also indicate a role in C4 photosynthesis, a debated topic in diatom research (Reinfelder et al., 2000; Johnston et al., 2001). (Discussed in chapter 4.)

Figure 3-23:
Sequence Alignment Overview for Contig 3 (2158bp, 14 clones, 20 sequences):

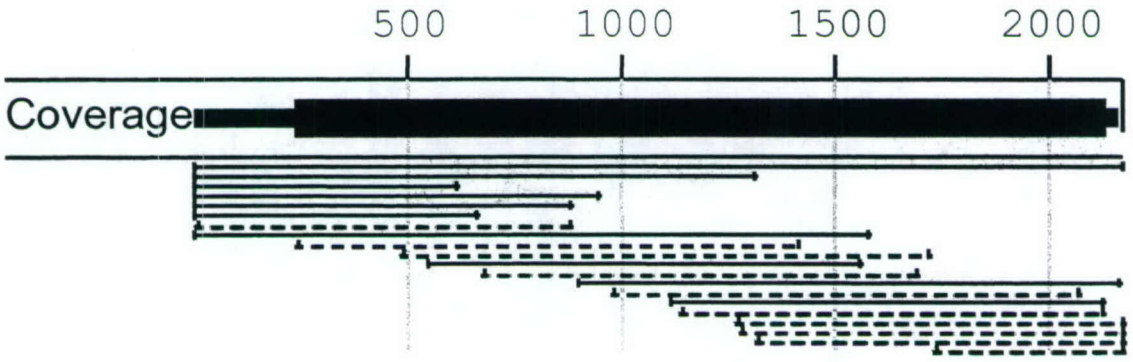


Figure 24: Predicted coding region for Contig 3:

Length: **532 aa**

Frame	from	to	Length
-2	102..	1700	1599

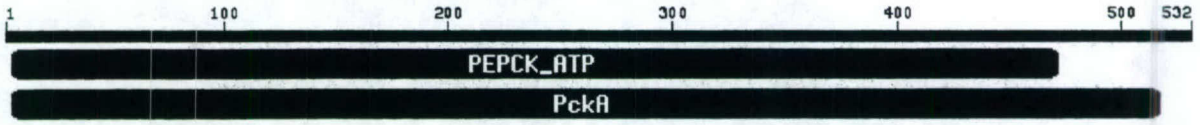
MTYQELFEHEQANNEGIVAKAKYGDTFRVSTGKYTGRSPKDKFVVLNPGSESAENMDWNDINQP
TSPEVYDELEDKAIKYFNTLDKAYVFDCYVGASPTSRKKIRFIHEMAWQQHFCTNMFIRPVSPEEL
DNFEPDFTVINACADEVEDHERLGLNSETAVVFNIEKGKGVIFGTWYGGENKKGIFSLMNYLLPLS
DPPQLPMHCSANVGKDHDVCLFFGLSGTGKTTLSADPHRALIGDDEHGWDEHGVYNFEGGCYAK
TINLSEETEPDIYRAIHTDALLENVMIESNTNVPNYFDTSITENGRVSYPFSIFSILIWSYHKEQMAGH
PKNIIFLSCDAFGVLPPVAKLSSGEAMYHFLSGYTAKVAGTERGIKEPVATFSTCFGAAFMTLHPTV
YADLLQKKLNTHGTNCYLVNSGWAGGPFGEGERMSIKTTRTCIDAILDGSIEDSKFENDPNFGFQV
PVALNGVSPEVLDIRSTWSDPAKYDEQAKKLGQMYIENFKKYEGKGSIDYTKFGPKNFGPDKEPK
SIFDL

Table 3-6: Fold-change Measurements for Individual cDNAs within Contig 3:

	cDNA Identifier	125D replicate spot 1	125D replicate spot 2	125D Average	S.D.	125C replicate spot 1	125C replicate spot 2	125C Average	S.D.	AX1 replicate spot 1	AX1 replicate spot 2	AX1 Average	S.D.
1	160E2	4.05	2.85	3.45	0.85	3.54	3.50	3.52	0.02	3.68	3.57	3.62	0.08
2	165A3	3.40	3.47	3.44	0.05	3.26	3.19	3.22	0.05	3.49	3.54	3.52	0.04
3	166F6	4.05	4.19	4.12	0.10	2.61	2.66	2.64	0.04	2.60	2.55	2.57	0.04
4	171B10	3.91	3.87	3.89	0.03	2.09	**	2.09	----	2.46	2.34	2.40	0.09
5	174F8	3.51	3.65	3.58	0.10	2.92	2.86	2.89	0.04	3.07	2.92	3.00	0.11
6	184H1	3.62	3.66	3.64	0.03	2.62	2.61	2.62	0.01	2.76	2.68	2.72	0.06
7	186G8	3.84	3.80	3.82	0.03	3.37	3.35	3.36	0.01	3.21	3.34	3.27	0.09
8	187D6	3.65	3.74	3.69	0.07	2.60	2.57	2.58	0.02	2.92	3.05	2.98	0.09
9	25D11	4.09	4.13	4.11	0.03	3.00	3.03	3.01	0.03	3.11	3.13	3.12	0.02
10	50C2	3.97	4.02	4.00	0.04	3.06	3.33	3.19	0.19	3.09	3.05	3.07	0.03
11	51D11	3.49	3.53	3.51	0.03	2.60	2.54	2.57	0.05	2.18	2.00	2.09	0.13
12	51H7	3.64	3.66	3.65	0.01	3.13	3.14	3.13	0.01	3.56	3.50	3.53	0.04
13	78C4	4.15	4.21	4.18	0.04	3.51	3.24	3.37	0.19	3.07	1.86	2.46	0.86
	AVERAGE	3.80	3.75	**3.78**		2.95	3.00	**2.94**		3.01	2.89	**2.95**	
	S.D.	0.26	0.36	**0.26**		0.43	0.34	**0.42**		0.43	0.57	**0.48**	

**Not significant

Figure 3-25: Contig 3 Sequence Alignment with COG2706, Conserved Domain for Phosphoenolpyruvate carboxykinase:



● gnl|CDD|4483, pfam01293, PEPCK_ATP, Phosphoenolpyruvate carboxykinase.

CD-Length = 471 residues, 97.9% aligned
Score = 602 bits (1553), Expect = 4e-173

```
Query:  3    YQELFEHEQANNEGIVAKAKYGDTFRVSTGKYTGRSPKDKFVVLNPGSESAENMDWND-I  61
Sbjct:  11   AAQLYEEALKNEKEGVLTST--GALAAMTGAKTGRSPKDKFIVDDETTR--DNIWWGSNV  66

Query:  62   NQPTSPEVYDELEDKAIKYFNTLDKAYVFDCYVGASPTSRKKIRFIHEMAWQQHFCTNMF  121
Sbjct:  67   NKPISEETFEILRERALDYLSTRDKLFVVDAFAGADPDYRLKVRVVTERAWHALFMRNML  126

Query:  122  IRPVSPEELDNFEPDFTVINACADEVEDHERLGLNSETAVVFNIEKGKGVIFGTWYGGEN  181
Sbjct:  127  IRPTPEEELTFFEPDFTIYNAGQFKA-DPKTHGLTSETFVAINFKRREQVILGTEYAGEM  185

Query:  182  KKGIFSLMNYLLPLSDPPQLPMHCSANVGKDHDVCLFFGLSGTGKTTLSADPHRALIGDD  241
Sbjct:  186  KKGIFSVMNYLLPEKG--ILSMHCSANVGKQGDVALFFGLSGTGKTTLSADPHRKLIGDD  243

Query:  242  EHGWDEHGVYNFEGGCYAKTINLSEETEPDIYRAIHTDALLENVMIESNTNVPNYFDTSI  301
Sbjct:  244  EHGWSDNGVFNIEGGCYAKCINLSAEKEPEIFNAIKFGAVLENVVLDEDTREVDFDDKSI  303

Query:  302  TENGRVSYP-FSIFSILIWSYHKEQMAGHPKNIIFLSCDAFGVLPPVAKLSSGEAMYHFL  360
Sbjct:  304  TENTRVAYPIEHIPNAVKTS-----KAPHPKNVIFLTADAFGVLPPVSKLTPEQAMYHFL  358

Query:  361  SGYTAKVAGTERGIKEPVATFSTCFGAAFMTLHPTVYADLLQKKLNTHGTNCYLVNSGWA  420
Sbjct:  359  SGYTAKLAGTERGVTEPEPTFSTCFGAPFLPLHPTKYAEMLAEKMAKHGADAWLVNTGWT  418

Query:  421  GGPFGEGERMSIKTTRTCIDAILDGSIEDSKFENDPNFGFQVPVALNGVSPEV  473
Sbjct:  419  GGSYGTGKRIPLKYTRAIIDAIHDGSLDNAEYETDPVFGLAIPTELPGVPSHI  471
```

● gnl|CDD|11576, COG1866, PckA, Phosphoenolpyruvate carboxykinase (ATP) [Energy production and conversion]

CD-Length = 529 residues, 94.7% aligned
Score = 652 bits (1684), Expect = 0.0

```
Query:  3    YQELFEHEQANNEGIVAKAKYGDTFRVSTGKYTGRSPKDKFVVLNPGSESAENMDWNDIN  62
Sbjct:  27   AAQLYEEAIRRGEGVLTAT---GALRVDTGIYTGRSPKDKFIVRDD--STRDTIWWGTRN  81

Query:  63   QPTSPEVYDELEDKAIKYFNTLDKAYVFDCYVGASPTSRKKIRFIHEMAWQQHFCTNMFI  122
Sbjct:  82   KPISPETFDRLKGDVTDYLSGKD-LFVVDGFAGADPDYRLPVRVVTEVAWHALFIRNLFI  140

Query:  123  RPVSPEELDNFEPDFTVINACADEVEDHERLGLNSETAVVFNIEKGKGVIFGTWYGGENK  182
Sbjct:  141  RP-TGEELSTFKPDFTVINAPSFK-ADPKRDGLRSETFVAFNFTERIVLIGGTWYAGEMK  198

Query:  183  KGIFSLMNYLLPLSDPPQLPMHCSANVGKDHDVCLFFGLSGTGKTTLSADPHRALIGDDE  242
Sbjct:  199  KGIFSVMNYLLPLKG--ILSMHCSANVGEKGDVALFFGLSGTGKTTLSADPHRRLIGDDE  256

Query:  243  HGWDEHGVYNFEGGCYAKTINLSEETEPDIYRAIHTDALLENVMIESNTNVPNYFDTSIT  302
Sbjct:  257  HGWDDRGVFNFEGGCYAKTINLSEEKEPEIYAAIKRGAVLENVVVDED-GTPDFDDGSLT  315

Query:  303  ENGRVSYPFSIFSILIWSYHKEQMAGHPKNIIFLSCDAFGVLPPVAKLSSGEAMYHFLSG  362
Sbjct:  316  ENTRAAYP--IEHIPNVS--PSVKAGHPKNVIFLTADAFGVLPPVSRLTPEQAMYHFLSG  371

Query:  363  YTAKVAGTERGIKEPVATFSTCFGAAFMTLHPTVYADLLQKKLNTHGTNCYLVNSGWAGG  422
Sbjct:  372  YTAKLAGTERGVTEPEPTFSTCFGAPFMPLHPTRYAELLGKLIKAHGANVYLVNTGWTGG  431

Query:  423  PFGEGERMSIKTTRTCIDAILDGSIEDSKFENDPNFGFQVPVALNGVSPEVLDIRSTWSD  482
Sbjct:  432  AYGTGKRIPIKYTRALLDAILDGSLENAETKTDPIFGLAIPVALPGVDSDILNPRNTWAD  491

Query:  483  PAKYDEQAKKLGQMYIENFKKYEGKGSIDYTKFGPK  518
Sbjct:  492  KAAYDEKARRLAKLFIENFKKYEDLADGAALVAAPP  527
```

Table 3-7: Contig 3, Blast Results, Overview:

PSN Identifier	NCBI Database	Consensus length (bp)	NCBI Identifier	Putative Identification	*Species* or Domain Name	E-value
Contig 3	NR	2158	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	e-164
Contig 3	CDD	2158	gnl_CDD_11576	phosphoenolpyruvate carboxykinase (ATP)	COG1866	0

Table 3-8: Contig 3, Individual cDNA BlastX Results, Overview:

	PSN Identifier	Length (bp)	NCBI Identifier	Putative Identification	*Species* or Domain Name	E-value
1	160E2	904	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	2.00E-13
1	165A3	1381	NP_710432.1	Phosphoenolpyruvate carboxykinase	Leptospira interrogans	5.00E-97
2	166F6	934	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	3.00E-81
3	16A12	781	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	4.00E-71
4	171B10	1320	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	3.00E-87
5	174F8	963	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	5.00E-72
6	184H1	963	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	1.00E-46
7	186G8	1314	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	e-127
8	187D6	941	O09460	phosphoenolpyruvate carboxykinase (ATP)	Anaerobiospirillum succiniciproducens	6.00E-24
9	25D11	977	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	e-113
10	50	908	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	7.00E-75
11	51D11	924	NP_282084.1	phosphoenolpyruvate carboxykinase (ATP)	Campylobacter jejuni	2.00E-46
12	51H7	845	ZP_00005661.2	phosphoenolpyruvate carboxykinase (ATP)	Rhodobacter sphaeroides	1.00E-63
13	78C4	869	NP_710432.1	Phosphoenolpyruvate carboxykinase	Leptospira interrogans	7.00E-73

Figure 3-26: Contig 3 Sequence Alignment with Phosphenolpyruvate Carboxykinase Sequences from *Arabidopsis thaliana (62% similarity, 45% identity), Saccharomyces cerevisiae (62%, 46%), COG1866 (66%, 55%), Escherichia coli (67%, 51%), Campylobacter jejuni (70%, 58%),and genome sequence from Thalassiosira pseudonana (77%, 73%)*

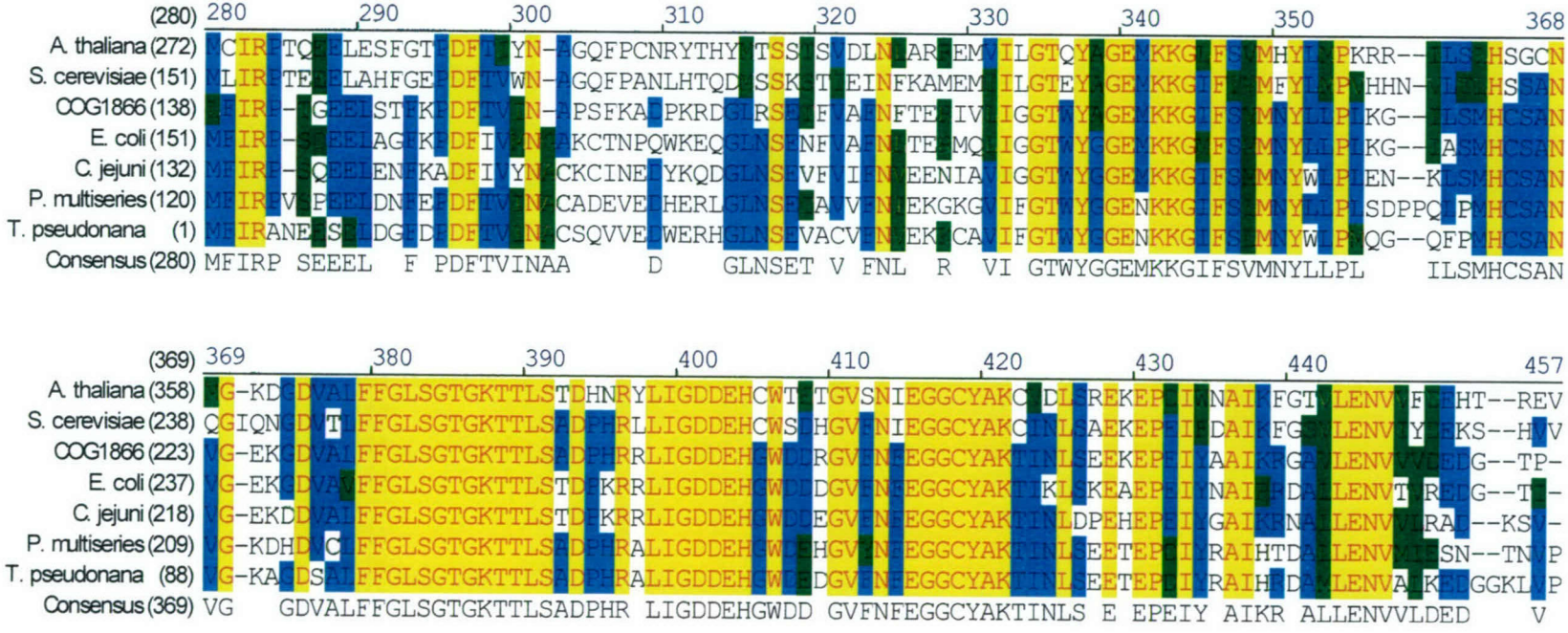


Figure 3-26: Contig 3 Sequence Alignment with Phosphenolpyruvate Carboxykinase Sequences from *Arabidopsis thaliana (62% similarity, 45% identity), Saccharomyces cerevisiae (62%, 46%), COG1866 (66%, 55%), Escherichia coli (67%, 51%), Campylobacter jejuni (70%, 58%),and genome sequence from Thalassiosira pseudonana (77%, 73%)*

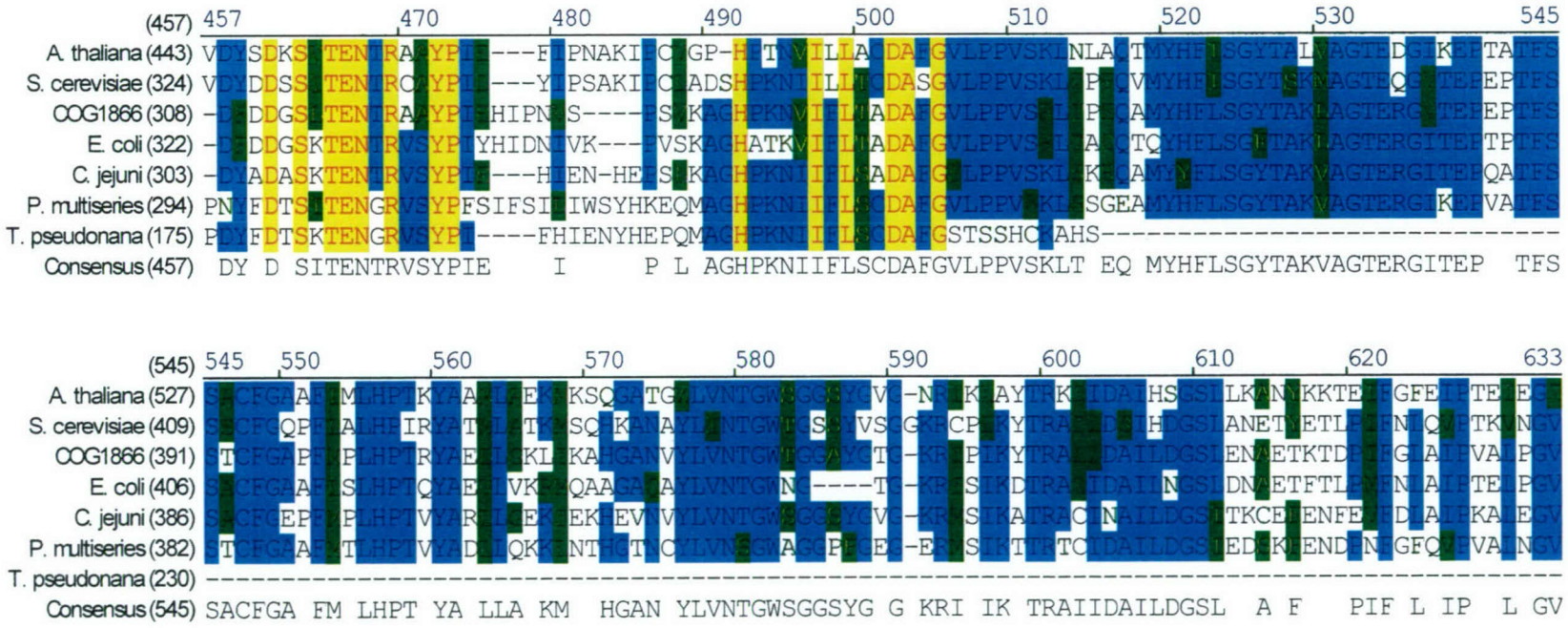
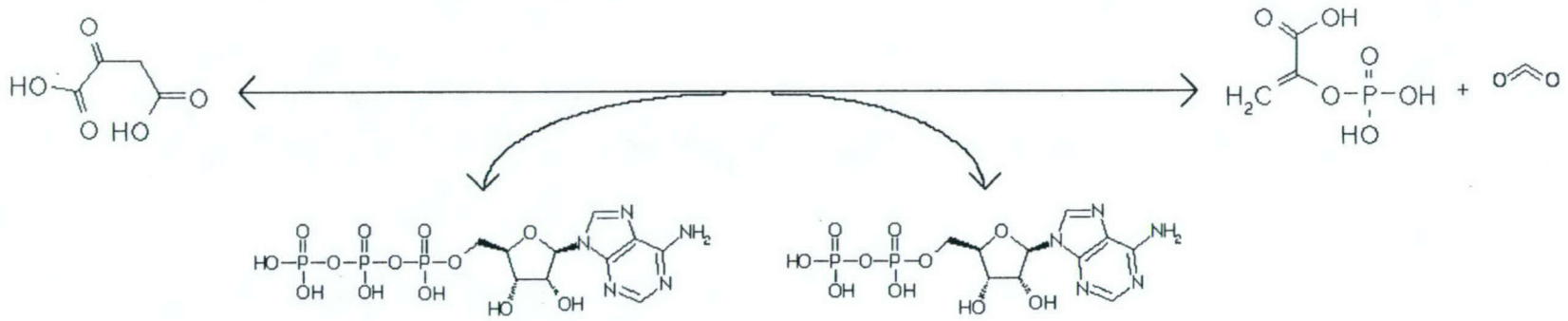


Figure 3-27: Phosphoenolpyruvate carboxykinase (ATP) - Reaction catalyzed:

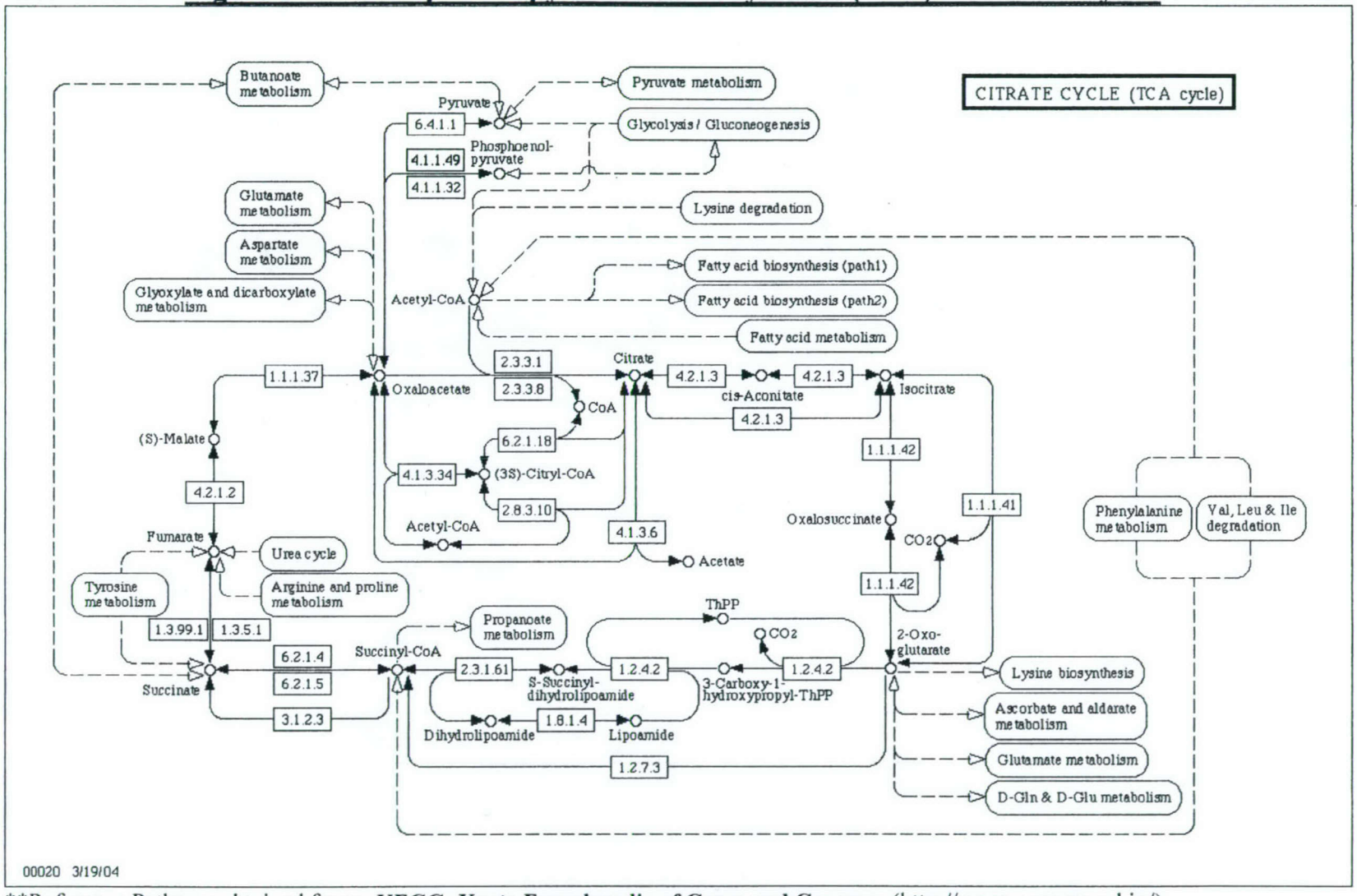
Oxaloacetate **Phosphoenolpyruvate + CO_2**



Adenosine 5'-triphosphate **Adenosine 5'-diphosphate**

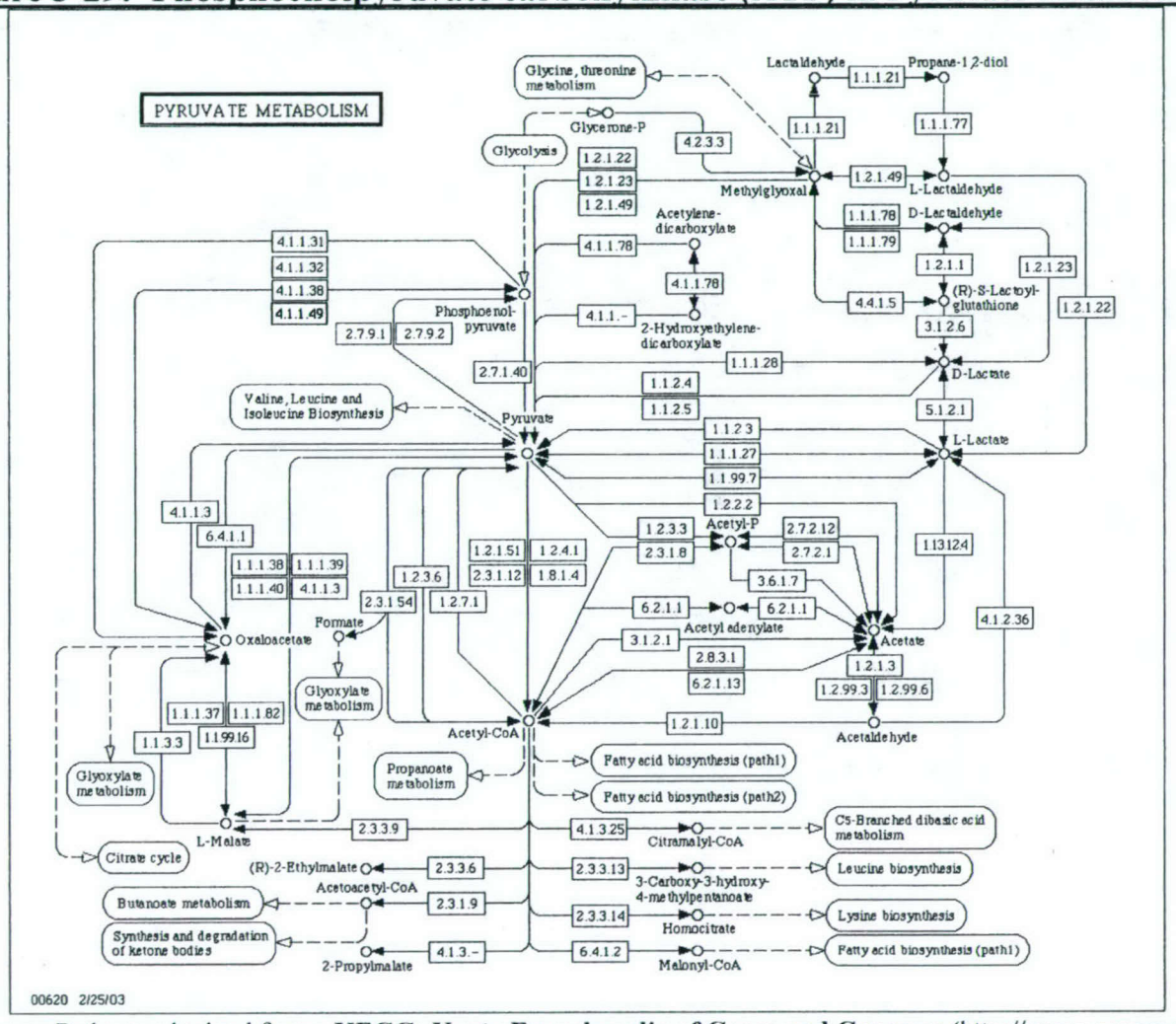
Reference Reaction obtained from: **KEGG: Kyoto Encyclopedia of Genes and Genomes http://www.genome.ad.jp/.

Figure 3-28: Phosphoenolpyruvate carboxykinase (ATP) in Citrate Cycle:



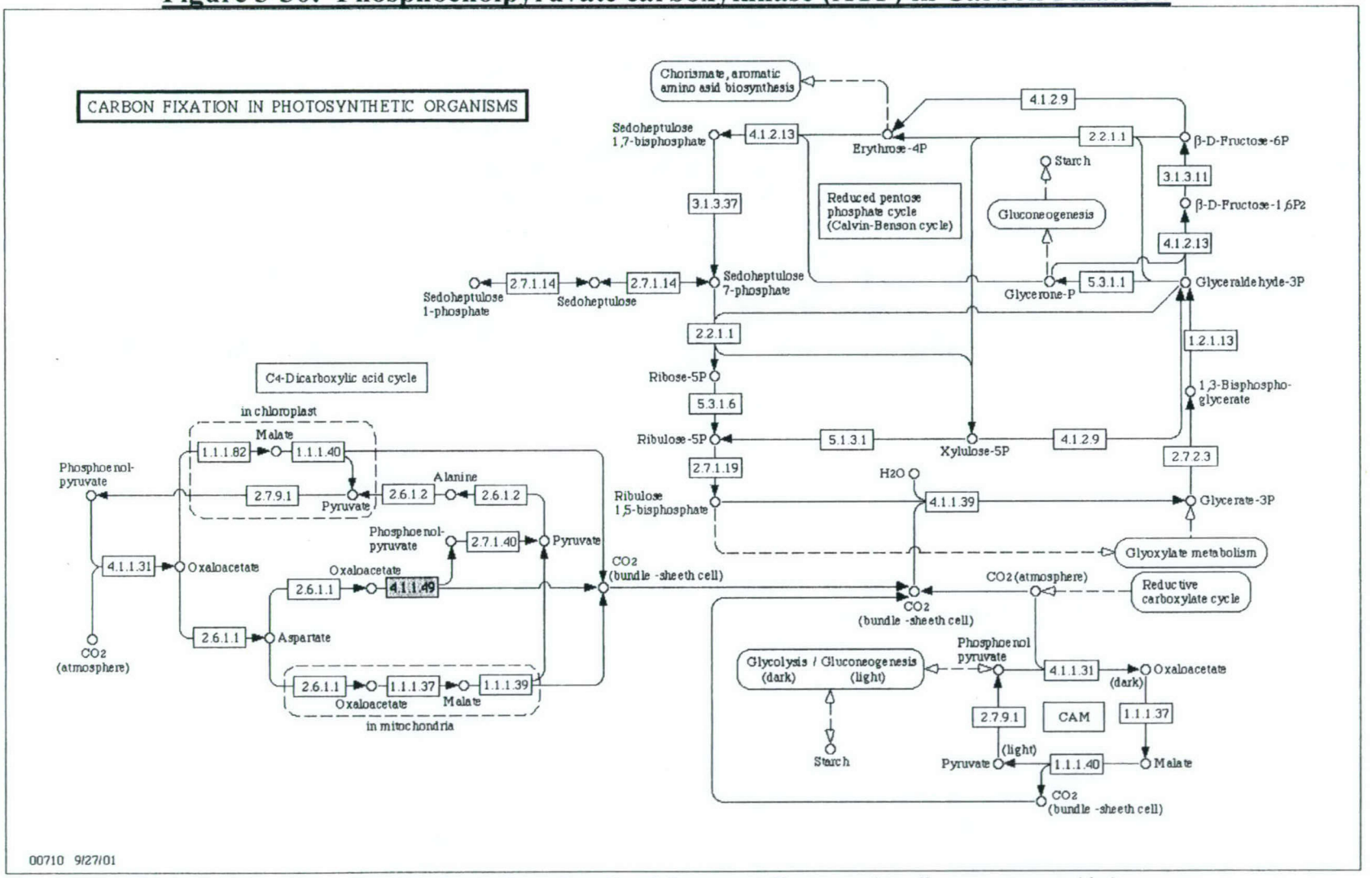
Reference Pathway obtained from: **KEGG: Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/)

Figure 3-29: Phosphoenolpyruvate carboxykinase (ATP) in Pyruvate metabolism:



Reference Pathway obtained from: **KEGG: Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/).

Figure 3-30: Phosphoenolpyruvate carboxykinase (ATP) in Carbon Fixation:



Reference Pathway obtained from: **KEGG: Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/).

Contig 4, Amino acid transporter: Contig 4 includes 3 cDNAs, which align to form a consensus sequence 1818 bp long (Figure 3-31). Overall average expression ratios were 3.24 (± 0.25) in Experiment 125C, 3.31 (± 0.0.7) in Experiment 125D, and 1.91 (± 0.23) in Experiment AX1 (Table 3-9). The predicted coding region for Contig 4 revealed an open reading frame (ORF) of 363 amino acids, which aligned with pfam00209, a sodium:neurotransmitter symporter family (Figures 3-32 and 3-33). The *P. multiseries* sequence aligned most closely with a novel human amino acid transporter, $hATB^{0+}$, with an E-value of 8E-34 (Table 3-10, 3-11, Figure 3-34). $hATB^{0+}$ is NA^{+}/Cl^{-} dependent member of the neurotransmitter symporter family, with the highest sequence similarity to the glycine and proline transporters. $hATB^{0+}$ was found to transport both neutral and cationic amino acids (Sloan and Mager, 1999.) Searching the *T. pseudonana* genome and *P. tricornutum* EST databases did not reveal any homologous sequences, suggesting the hypothesis that this amino acid transporter is unique to *Pseudo-nitzschia* spp and not present in non-toxin-producing diatoms. If this transporter is unique to *P. multiseries*, it may be that the transporter is actively involved in export of DA from the cell or imports a precursor to DAinto the cell.

Figure 3-31: Sequence Alignment Overview for Contig 4 (1818bp, 3 clone consensus sequences):

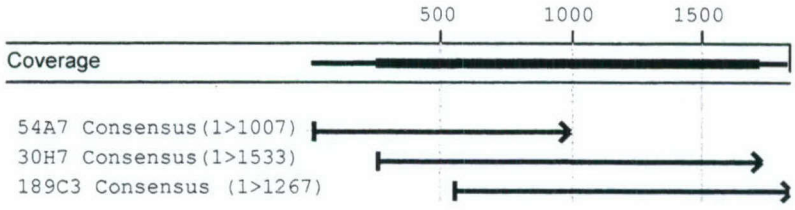


Figure 3-32: Predicted coding region for Contig 4:

Length: **363 aa**

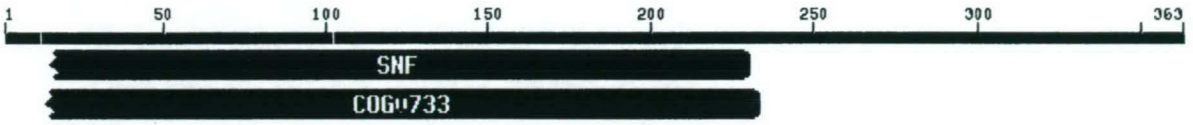
Frame	from	to	Length
+1	49..	1140	1092

MGLPVVLLFVFLGKALTLEGSSDGVKEYIGVWDVSVLTTRGEVWSVACSQIFFSISLTFGILTAFGS
HCPRSEPAVANAVVVSLSNSLFSFVSGFAVFAALGHLAFLENKEPTDLEFKGFGLVFGTWPVVFNT
LPGGIHWVRLIFFNLFLLGIDSAFAFLEAFITVMHDTVYFEKIPRFWLAAGISMVGFLFSLMYCSDA
GLFWLDVIDFYINFVMILVGFFEAFGSAWAYDLPGQIERQTAPVVYSFWTANFGAVALGCILWFST
NPDVAVWAGFVGFFGWYFAFVGVTHFYITKALANDTENKWTVKSLWWEVYFGKHSVPSRPHAR
SYWKGPFCLVSVDEAFHPTRPYHSLCQLGAIEER

Figure 3-9: Fold-change Measurements for Individual cDNAs within Contig 4:

	cDNA Identifier	125D replicate spot 1	125D replicate spot 2	125D Average	S.D.	125C replicate spot 1	125C replicate spot 2	125C Average	S.D.	AX1 replicate spot 1	AX1 replicate spot 2	AX1 Average	S.D.
1	189C3	3.33	3.23	3.28	0.07	3.38	3.26	3.32	0.09	2.14	1.92	2.03	0.16
2	30H7	4.25	4.19	4.22	0.04	3.49	3.60	3.54	0.08	2.13	2.10	2.12	0.02
3	54A7 rep 1	2.56	2.46	2.51	0.07	2.90	3.30	3.10	0.28	1.56	1.64	1.60	0.06
	54A7 rep 2	3.16	3.31	3.23	0.11	2.90	3.07	2.98	0.12	1.82	2.00	1.91	0.13
	AVERAGE	3.33	3.30	**3.31**		3.17	3.31	**3.24**		1.91	1.92	**1.91**	
	S.D.	0.70	0.71	**0.70**		0.31	0.22	**0.25**		0.28	0.20	**0.23**	

Figure 3-33: Contig 4 Sequence Alignment with pfam00209, COG0733, Conserved Domain for sodium:neurotransmitter:



gnl|CDD|25443, pfam00209, SNF, Sodium:neurotransmitter symporter family.

Conserved Domain -Length = 536 residues, 40.7% aligned
Score = 94.5 bits (235), Expect = 2e-20

```
Query:  14   KALTLEGSSDGVKEYIGVWDVSVLTTRGEVWSVACSQIFFSISLTFGILTAFGSHCPRSE  73
Sbjct:  221  RGVTLPGAADGIDFYLTP-DWSKLL-DPQVWIDAATQIFFSLGIGFGGLIAFASYNKFHN  278

Query:  74   PAVANAVVVSLSNSLFSFVSGFAVFAALGHLAFLENKEPTDLEFKGFGLVFGTWPVVFNT  133
Sbjct:  279  NCYRDALIVSFINSATSFLAGFVIFSVLGFMAQEQGVPISEVAESGPGLAFIAYPEAVTM  338

Query:  134  LPGGIHWVRLIFFNLFLLGIDSAFAFLEAFITVMHDTV-YFEKIPRFWLAAGISMVGFLF  192
Sbjct:  339  LPLSPFWSVLFFLMLITLGLDSQFGGVEGIITALVDEFPIVLRVRREVFTLGVCVISFLI  398

Query:  193  SLMYCSDAGLFWLDVIDFYINFVMIL-VGFFEAFGSAWAY  231
Sbjct:  399  GLLFVTEGGIYVFTLFDYYAASFGLLFVAFFECIAIAWVY  438
```

gnl|CDD|10602, COG0733, COG0733, Na+-dependent transporters of the SNF family

Conserved Domain -Length = 439 residues, 47.8% aligned
Score = 68.3 bits (167), Expect = 1e-12

```
Query:  13   GKALTLEGSSDGVKEYIGVWDVSVLTTRGEVWSVACSQIFFSISLTFGILTAFGSHCPRS  72
Sbjct:  187  IRAVTLPGAMEGLK-FLFKPDFSKLT-DPKVWLAALGQAFFSLSLGFGIMITYSSYLSKK  244

Query:  73   EPAVANAVVVSLSNSLFSFVSGFAVFAALGHLAFLENKEPTDLEFKGFGLVFGTWPVVFN  132
Sbjct:  245  SDLVSSALSIVLLNTLISLLAGLVIFPALFSFGADAS--------QGPGLVFIVLPAVFN  296

Query:  133  TLPGGIHWVRLIFFNLFLLGIDSAFAFLEAFITVMHDTVYFEKIPRFWLAAGISMVGFLF  192
Sbjct:  297  QMPLGTLFGILFFLLLLFAALTSAISMLEVLVAALIDKF---GISRKKATWLIGILIFLL  353

Query:  193  SLMYCSDAGLFWLDVIDFYI-NFVMILVGFFEAFGSAWAYDLP  234
Sbjct:  354  GIPSILSFGLSIFDLVDFVVSNILMPLGALLIAIFVGWVLKKE  396
```

Table 3-10: Contig 4, Blast Results, Overview:

PSN Identifier	NCBI Database	Consensus length (bp)	NCBI Identifier	Putative Identification	*Species* or Domain Name	E-value
Contig 4	CDD	1818	Gnl_CDD_25443	Sodium: neurotransmitter symporter family	Pfam00209	2e-20
Contig 4	NR	1818	NP_009162.1	solute carrier family 6 (neurotransmitter transporter), member 14 amino acid transporter B0+	*Homo sapiens*	8.00E-34

Table 3-11: Contig 4, Individual cDNA BlastX Results - NR, Overview:

	PSN Identifier	Length (bp)	NCBI Identifier	Putative Identification	*Species* or Domain Name	E-value
1	189C3	1267	NP_070819.1	sodium- and chloride-dependent transporter	*Archaeoglobus fulgidus*	5.00E-07
2	30H7	1533	NP_009162.1	solute carrier family 6 (neurotransmitter transporter), member 14 amino acid transporter B0+	*Homo sapiens*	1.00E-12
3	54A7	1007	NP_009162.1	solute carrier family 6 (neurotransmitter transporter), member 14 amino acid transporter B0+	*Homo sapiens*	3.00E-24

Figure 3-34: Contig 4 Sequence Alignment Against hATB^{O+}

>ref|NP_009162.1| solute carrier family 6 (neurotransmitter transporter), member 14 amino acid transporter B0+ [Homo sapiens]

Score = 146 bits (368), Expect = 8e-34
Identities = 84/232 (36%), Positives = 123/232 (53%), Gaps = 6/232 (2%)

```
Query: 4   PVVLLFVFLGKALTLEGSSDGVKEYIGVWDVSVLTTRGEVWSVACSQIFFSISLTFGILT 63
           P V+L + L +  TLEG+S G+  YIG           EVW  A +QIF+S+S+ +G L
Sbjct: 272 PYVVLLILLVRGATLEGASKGISYYIGAQSNFTKLKEAEVWKDAATQIFYSLSVAWGGLV 331

Query: 64  AFGSHCPRSEPAVANAVVVSLSNSLFSFVSGFAVFAALGHLAFLENKEPTDLEFKGFGLV 123
           A  S+         ++A+VV L+N L S  +GFA+F+ LGH+A +  KE + +   GF L
Sbjct: 332 ALSSYNKFKNNCFSDAIVVCLTNCLTSVFAGFAIFSILGHMAHISGKEVSQVVKSGFDLA 391

Query: 124 FGTWPVVFNTLPGGIHWVRLIFFNLFLLGIDSAFAFLEAFITVMHDTVYFEKI---PRFW 180
           F  +P     LPGG  W  L FF L  LG+DS FA +E   T + D   F K+    R
Sbjct: 392 FIAYPEALAQLPGGPFWSILFFFMLLTLGLDSQFASIETITTTIQD--LFPKVMKKMRVP 449

Query: 181 LAAGISMVGFLFSLMYCSDAGLFWLDVID-FYINFVMILVGFFEAFGSAWAY 231
           +  G  +V FL  L+  + AG++W+ +ID F   + +++    E  G  W Y
Sbjct: 450 ITLGCCLVLFLLGLVCVTQAGIYWVHLIDHFCAGWGILIAAILELVGIIWIY 501
```

NO HITS against *Thalassiosira pseudonana

Contig 5, Small Heat Shock Protein: Contig 5 includes seven cDNAs, which form a consensus sequence 1001 bp long (Figure 3-35). Overall expression ratios were very high in all three experiments, which averaged 7.00 (± 0.16) in Experiment 125C, 7.81 (± 0.43) in Experiment 125D, and 4.40 (± 0.41) in Experiment AX1 (Table 3-12). The predicted open reading frame of 209aa appears to be a small heat shock protein (Figure 3-36, Table 3-13, 3-14). Pfam HMM analysis revealed homology with hsp20, a family of alpha-crystallin hsps (E-value – 9.7E-23)(Figure 3-37). The alpha-crystallin-type heat shock proteins are a family of small stress-induced proteins ranging from 12 to 43 kDa, whose common feature is the alpha-crystallin domain. Generally active as large oligomers consisting of multiple subunits, these proteins are believed to be ATP-independent chaperones that prevent stress-induced denaturation and aggregation, and are important in refolding in combination with other heat shock proteins (Narberhaus, 2002). The induction of a small heat shock protein is consistent with the conditions of cell growth, given that cells would be stressed during stationary growth as certain environmental conditions become limiting. Interestingly, similarity searching against the *T. pseudonana* database did not reveal any homology and searching against the *P. tricornutum* database revealed a short sequence of 39 bp with weak similarity (14/39 [35%] identity). Follow-up expression studies to determine functionality will shed light on whether this is truly a stress response protein or if it may function in DAmetabolism.

Figure 3-35: Sequence Alignment Overview for Contig 5 (1001bp, 7clones):

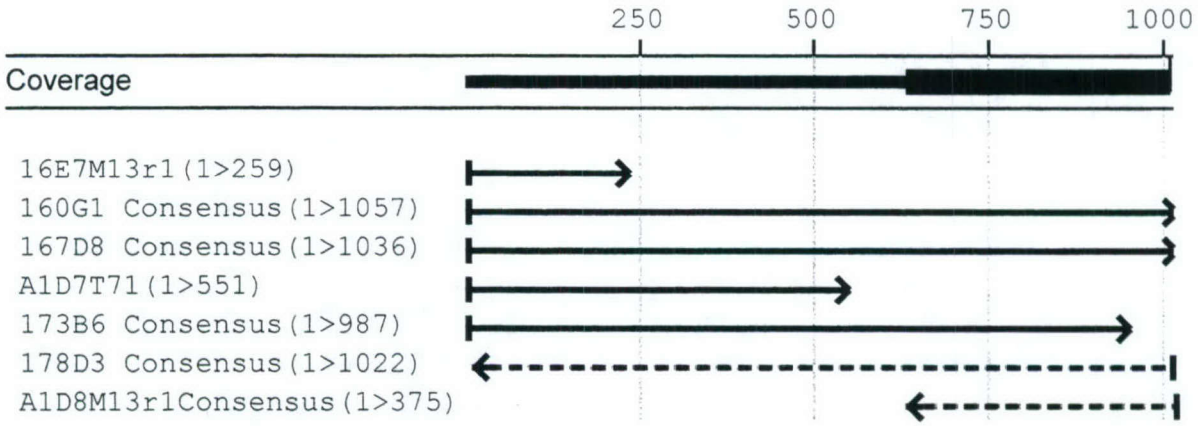


Figure 3-36: Predicted coding region for Contig 5:

Length: **209 aa**

Frame	from	to	Length
+2	53..	682	630

MSMNLSKQAAGALAFAPFASPWSWGFGPSHFYSLAPSMRLAESAFDTDIERAIEKQQKLAQRMW
DQASVVPQVLGQHNRYELIDNDEKFQLTVDVPGIKQDDIDIKLDEGFITVEGHREATTKNSRFSSKF
AQTFSLDPAVDVKKITATLDNGVLVVAAPKEAAKLEEKPRRIPIQTMKKKEELKAAKHDIPVETVG
EKEEVMDLDKEN

Table 3-12: Fold-change Measurements for Individuals cDNAs within Contig 5:

cDNA Identifier	125D replicate spot 1	125D replicate spot 2	125D Average	S.D.	125C replicate spot 1	125C replicate spot 2	125C Average	S.D.	AX1 replicate spot 1	AX1 replicate spot 2	AX1 Average	S.D.
178D3	7.33	6.76	7.05	0.40	8.05	7.53	7.79	0.37	4.67	4.55	4.61	0.09
173B6	7.05	7.21	7.13	0.11	7.45	8.76	8.11	0.93	4.71	4.65	4.68	0.04
167D8	7.05	7.06	7.05	0.00	7.29	7.10	7.20	0.13	3.71	3.86	3.78	0.10
160G1	6.85	6.68	6.76	0.12	7.43	8.82	8.12	0.99	4.54	4.48	4.51	0.04
AVERAGE	7.07	6.93	**7.00**		7.56	8.05	**7.81**		4.41	4.38	**4.40**	
S.D.	0.20	0.25	**0.16**		0.34	0.87	**0.43**		0.47	0.36	**0.41**	

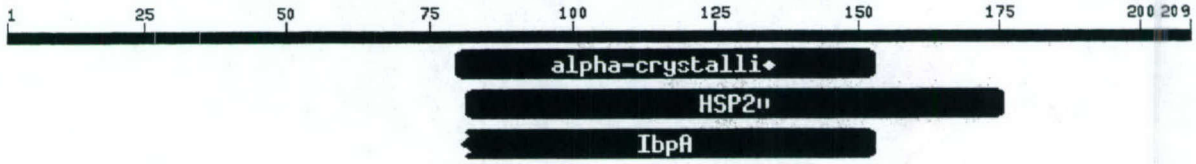
Table 3-13: Contig 5 Blast Results, Overview:

PSN Identifier	Database	Consensus Sequence Length (bp)	Identifier	Description	*Species* or Domain Name	E-value
Contig 5	Pfam	1001	HMM HSP20	Hsp20/alpha crystallin	HSP 20 Seed	9.7e-23
Contig 5	NR	1001	gnl|CDD|16554	alpha-crystallin-Hsps	cd00298	3e-11
Contig 5	Pfam	1001	HMM HSP20	Hsp20/alpha crystallin	HSP 20 Seed	9.7e-23

Table 3-14: Contig 5, Individual cDNA BlastX Results - NR, Overview:

	PSN Identifier	Length (bp)	NCBI Identifier	Putative Identification	*Species* or Domain Name	E-value
1	160G1	1057	NP_103744.1	small heat shock protein	Mesorhizobium loti	2.00E-08
2	167D8	1036	NP_103744.1	small heat shock protein	Mesorhizobium loti	2.00E-05
3	16E7	259		NO HITS		
4	173B6	987	ZP_00170792.2	small heat shock protein	Ralstonia eutropha	0.006
5	178D3	1022	NP_103744.1	small heat shock protein	Mesorhizobium loti	4.00E-07
6	A1D7	551	NP_103744.1	small heat shock protein	Mesorhizobium loti	8.00E-10
7	A1D8	375				>.01

Figure 3-37: Contig 5 Sequence Alignment with pfam00011, cd00298, COG0071, Conserved Domain for small Heat Shock Proteins:



● gnl|CDD|16554, cd00298, alpha-crystallin-Hsps, alpha-crystallin-type heat shock proteins (Hsps)

CD-Length = 89 residues, 89.9% aligned
Score = 62.9 bits (153), Expect = 3e-11

```
Query:  79   RYELIDNDEKFQLTVDVPGIKQDDIDIKLDEGFITVEGHREATTKNS-----RFSSKFAQ  133
Sbjct:  2    PVDIKEDDEHFEVKLDVPGFKPEDLKVKVEDNVLVVKGKREEEQDEKSLRHGRSSREFSR  61

Query:  134  TFSLDPAVDVKKITATLDNG  153
Sbjct:  62   KFTLPENVDPDAIKASLSNG  81
```

● gnl|CDD|25360, pfam00011, HSP20, Hsp20/alpha crystallin family.

CD-Length = 102 residues, **100.0% aligned**
Score = 60.6 bits (147), Expect = 1e-10

```
Query:  81   ELIDNDEKFQLTVDVPGIKQDDIDIKLDEGFITVEGHREATTKNS------RFSSKFAQT  134
Sbjct:  1    DIKEDKDAFVVKLDVPGFKPEELKVKVEDNRVLVKGKHEEEEEDDHGLRSERSYGSFSRK  60

Query:  135  FSLDPAVDVKKITATLDNGVLVVAAPKEAAKLEEKPRRIPIQ  176
Sbjct:  61   FTLPENADPDKVKASLKNGVLTVTVPKLEPEEDKKERRIQIQ  102
```

● gnl|CDD|9946, COG0071, IbpA, Molecular chaperone (small heat shock protein) [Posttranslational modification, protein turnover, chaperones]

CD-Length = 146 residues, 56.2% aligned
Score = 60.8 bits (147), Expect = 1e-10

```
Query:  80   YELIDNDEKFQLTVDVPGIKQDDIDIKLDEGFITVEGHREATTKNS--------RFSSKF  131
Sbjct:  43   VDIEETDDEYRITAELPGVDKEDIEITVEGNTLTIRGEREEEEEEEEEGYLRRERAYGEF  102

Query:  132  AQTFSLDPAVDVKKITATLDNG  153
Sbjct:  103  ERTFRLPEKVDPEVIKAKYKNG  124
```

Contig 6, Acyl-CoA synthetase (AMP-forming): Contig 6 includes 13 cDNAs, which align to form a consensus sequence 2438 bp long (Figure 3-38). Overall average expression ratios were 4.66 (± 0.92) in Experiment 125C, 3.76 (± 0.79) in Experiment 125D, and 2.13 (± 0.26) in Experiment AX1 (Table 3-15). Contig 6 revealed a coding region that was split into four separate reading frames. Each deduced sequence showed high homology to AMP-forming acyl-coA synthetase, so the coding regions were spliced together for subsequent analysis (Figure 3-39). The deduced protein aligned closely with this family of enzymes that act via an ATP-dependent covalent binding of AMP to their substrate (Figure 3-40, Table 3-16); these enzymes have been shown to function in lipid metabolism, secondary metabolite biosynthesis, transport, and catabolism (Faergeman et al., 1997; Sharma et al., 1996; Black et al., 1992). Up-regulation of this transcript evokes the suggestion that it may function in the formation of the isoprenoid side chain of DA. *T. pseudonana* alignment produced multiple hits, with two areas of high sequence identity on scaffold 1 (Figure 3-41).

Figure 3-27: Sequence alignment Overview for Contig 6 (2438bp, 13 clones):

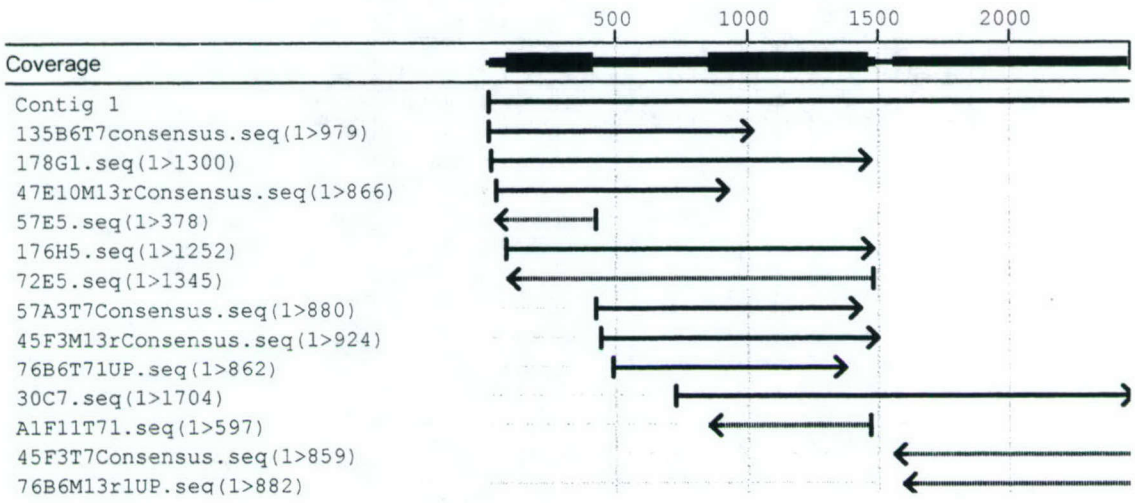
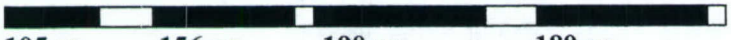


Figure 3-28: Predicted coding region for Contig 6:



105 aa **156 aa** **190 aa** **189 aa**

Frame		from	to	Length
+3	●	1038..	1610	573
+3	●	1788..	2357	570
+3	●	498..	968	471
+3	●	3..	320	318

WXXXHCAGSYSNQKETIDDTSILFRTKKLNIFYADNSNQKRGD
GRWRFFNVTSLKGARRYLFPTGTKRTVYSFFSTVESTSNANGFQ
QKDIQFKTLYELQKTACD

MIAAATYSLNATLVPMYEAQLATDWKYIINDSSASVIICSTKDI
FHRFSNEVLHMTPSVHSTLCLDAVDGEEYGYQTAMSEIIDNSLPPEQSSIVTTPNEED
LANLIYTSGTTGNPKGVELTHRNIVSNIKGGRLLSQKPHDLLDESSKTLAFLPW

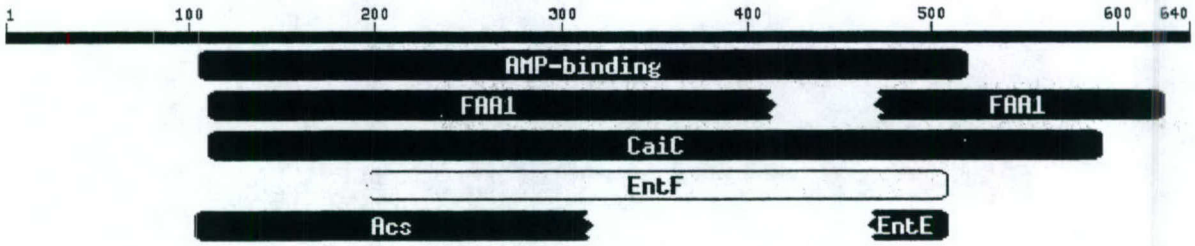
MNFISLTYRAHSYGQTVELWMAMSFGSSCAICRGIPFLLEDLQL
VKPTVIFAVPTLYSKIYDSVQNKANSGQYIQDALLRNAIEIGNKNAQFRRGERDALSF
AQTLKYTVLDRLVLSKIRDRFGGNLRYSCVAGAACPIDVLKFMDSIGIPVLEGLWSYR
DVTDHFAQFYQQTINRKRRTSTTRCRSLHH

MQEQSSRILLFISFFRFHTGDLGRMDADGWIKVTGRIKEQYKLE
NGKYVAPAPIENAIGFSRFINQVVICGANKPYNVALIVPDWPAISQKLEYNDEIPESE
IANKKQVHILIDNEIRKCCSDLKKFEIPRKWAFVAPFTAANNMVTPKMSIRKHKVLET
YSDLIANLYRNDTADSNHADDQYTFDEAA

Table 3-15: Fold-change Measurements for Individual cDNAs within Contig 6:

	cDNA Identifier	125D replicate spot 1	125D replicate spot 2	125D Average	S.D.	125C replicate spot 1	125C replicate spot 2	125C Average	S.D.	AX1 replicate spot 1	AX1 replicate spot 2	AX1 Average	S.D.
1	135B6	6.01	5.97	5.99	0.03	4.71	4.57	4.64	0.10	2.44	2.45	2.45	0.01
2	173H1	4.82	4.89	4.86	0.05	3.55	2.63	3.09	0.65	2.21	1.69	1.95	0.37
3	176H5	5.36	5.51	5.44	0.10	3.63	3.25	3.44	0.27	2.08	2.14	2.11	0.04
4	178G1	5.07	4.67	4.87	0.28	4.14	2.77	3.46	0.97	2.11	2.10	2.10	0.00
5	180G9	4.50	4.19	4.35	0.22	3.30	3.44	3.37	0.10	1.97	2.03	2.00	0.04
6	182H3	3.13	3.29	3.21	0.11	2.79	2.68	2.73	0.08	2.04	1.96	2.00	0.05
7	30C7	5.37	3.39	4.38	1.40	3.57	4.56	4.06	0.70	2.40	**	2.40	-----
8	45F3	5.43	5.59	5.51	0.12	5.41	5.24	5.33	0.12	2.27	2.32	2.30	0.03
9	47E10	4.10	4.20	4.15	0.07	4.11	3.64	3.87	0.33	1.85	1.83	1.84	0.01
10	57A3	4.52	4.71	4.62	0.13	3.53	3.31	3.42	0.15	2.27	**	2.27	----
11	57E5	2.60	2.89	2.75	0.20	2.71	2.55	2.63	0.11	1.62	1.54	1.58	0.06
12	72E5	4.85	4.92	4.88	0.05	4.30	4.41	4.36	0.08	2.26	2.29	2.27	0.02
13	76B6	5.43	5.70	5.56	0.20	4.56	4.44	4.50	0.08	2.44	2.53	2.48	0.06
		4.71	4.61	**4.66**		3.87	3.65	**3.76**		2.15	2.08	**2.13**	
		0.96	0.98	**0.92**		0.77	0.90	**0.79**		0.24	0.31	**0.26**	

Figure 3-40: Contig 6 Sequence Alignment with Conserved Domain for Acyl-coA synthetase:



gnl|CDD|10750, COG1022, FAA1, Long-chain acyl-CoA synthetases (AMP-forming)

CD-Length = 613 residues
Score = 185 bits (471), Expect = 1e-47

```
Query:  110  ATYSLNATLVPMYEAQLATDWKYIINDSSASVIICSTKDIFHRFSNEVLHMTPSVH----  165
Sbjct:  89   AILALGAVSVPIYSTSTPEQLAYILNESESKVIFVENQELLDLV-LPVLEDCPKVVDLIV  147

Query:  166  -STLCLDAVDGEEYGYQTAMSEIID--NSLPPEQSSIVTTPNEEDLANLIYTSGTTGNPK  222
Sbjct:  148  IIDLVREAVEAKALVLEVFPDEGISLFLIDSAGLEGRIAPPKPDDLATIIYTSGTTGTPK  207

Query:  223  GVELTHRNIVSNIKGGRLLSQKPHDLLDESSKTLAFLPWMNFISLTYRAHSYGQTVELWM  282
Sbjct:  208  GVMLTHRNLLAQVAGIDEVLPPIG----PGDRVLSFLPL---------AHIFERAFEGGL  254

Query:  283  AMSFGSSCAICRGIPFLLEDLQLVKPTVIFAVPTLYSKIYDSVQNKANSGQYIQDALLRN  342
Sbjct:  255  ALYGGVTVLFKEDPRTLLEDLKEVRPTVMIGVPRVWEKVYKGIMEKVAKAPAVRRKLFRW  314

Query:  343  AIEIGNKNAQFRRGERDALSFAQTLKYTVLDRLVLSKIRDRFGGNLRYSCVAGAACPIDV  402
Sbjct:  315  ALKVAYK-----KISRALLGGGPLSWLLVADRLVFRKIRDALGGRIRYALSGGAPLSPEL  369

Query:  403  LKFMDSIGIPVLEG  416
Sbjct:  370  LHFFRSLGIPILEG  383
```

gnl|CDD|10192, COG0318, CaiC, Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II

CD-Length = 534 residues, 80.9% aligned
Score = 140 bits (353), Expect = 5e-34

```
Query:  110  ATYSLNATLVPMYEAQLATDWKYIINDSSASVIICSTKDIFHRFSNEVLHMTPSVHSTLC  169
Sbjct:  82   AALRAGAVAVPLNPRLTPRELAYILNDAGAKVLITSAE-----FAALLEAVAEALPVVLV  136

Query:  170  LDAVDGEEYGYQTAMSEIIDNSLPPEQSSIVTTPNEEDLANLIYTSGTTGNPKGVELTHR  229
Sbjct:  137  VLLVGDADDRLPITLEALAAEGPGPDADARPVDP--DDLAFLLYTSGTTGLPKGVVLTHR  194

Query:  230  NIVSNIKGGRLLSQKPHDLLDESSKTLAFLPWMNFISLTYRAHSYGQTVELWMAMSFGSS  289
Sbjct:  195  NLLANAAGIAAALG---GGLTPDDVVLSWLPLF---------HIFGLIVGLLAPLLGGGT  242

Query:  290  CAICRGIPF----LLEDLQLVKPTVIFAVPTLYSKIYDSVQNKANSGQYIQDALLRNAIE  345
Sbjct:  243  LVLLSPEPFDPEEVLWLIEKYKVTVLSGVPTFLR----------------ELLDNP--  282

Query:  346  IGNKNAQFRRGERDALSFAQTLKYTV-----LDRLVLSKIRDRFGGNLRY--------SC  392
Sbjct:  283  ----------EKDDDDLSSSLRLVLSGGAPLPPELLERFEERFGPIAILEGYGLTETSP  331

Query:  393  VAGAACPIDVLKFMDSIGIPVLEGLWSYRDVTDHFAQFYQQTINRKRRTSTTRCRSLHHM  452
Sbjct:  332  VVTINPPDDLLAKPGSVGRPLPGVEV--RIVDPDGGEVLPGEVGEIWVRGPNVMKGYWNR  389

Query:  453  QEQSSRILLFISFFRFHTGDLGRMDADGWIKVTGRIKEQYKLENGKYVAPAPIENAIGFS  512
Sbjct:  390  PEATAEAFDEDGWL--RTGDLGYVDEDGYLYIVGRLKDLIIS-GGENIYPEEIEAVLAEH  446

Query:  513  RFINQVVICGANKPYN----VALIVPDWPAISQKLEYNDEIPESEIANKKQVHILIDNEI  568
Sbjct:  447  PAVAEAAVVGVPDERWGERVVAVVVL-------KPGGDAELTAEEL-------------R  486

Query:  569  RKCCSDLKKFEIPRKWAFVA--PFTAA  593
Sbjct:  487  AFLRKRLALYKVPRIVVFVDELPRTAS  513
```

Table 3-16: Contig 6 Blast Results, Overview:

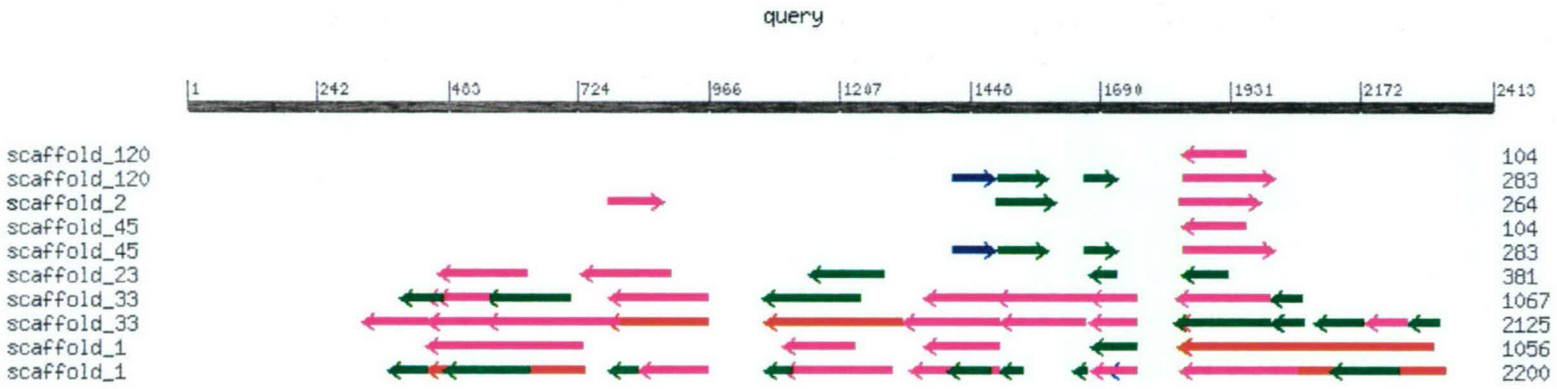
PSN Identifier	Database	Consensus Length (bp)	NCBI Identifier	Description	*Species* or Domain Name	E-value
Contig 6	CDD	2438	gnl|CDD|10750,	Long-chain acyl-CoA synthetases (AMP-forming)	COG1022	1.00E-47
Contig 6	NR	2438	NP_662047.1	long-chain-fatty-acid-CoA ligase	Chlorobium tepidum	4.00E-48

Table 3-17: Contig 6, Individual cDNA BlastX Results, Overview:

	PSN Identifier	Length (bp)	NCBI Identifier	Description	*Species* or Domain Name	E-value
1	135B6	979	NP_446059.1	fatty acid Coenzyme A ligase, long-chain fatty acid coenzyme A ligase 5	Rattus norvegicus	3.00E-20
2	173H1	1374	NP_662047.1	long-chain-fatty-acid-CoA ligase, putative	Chlorobium tepidum TLS	2.00E-31
3	176H5	1252	NP_446059.1	fatty acid Coenzyme A ligase, long chain 5; long-chain fatty acid	Rattus norvegicus	4.00E-30
4	178G1	1300	BAB16604.1	acyl-CoA synthetase 5	Cavia porcellus	5.00E-30
5	180G9	898	NP_960859.1	FadD15	Mycobacterium avium	6.00E-16
6	182H3	862	NP_960859.1	FadD15	Mycobacterium avium	3.00E-20
7	30C7	1704	NP_662047.1	long-chain-fatty-acid-CoA ligase, putative	Chlorobium tepidum	5.00E-44
8	45F3	924	NP_662047.1	long-chain-fatty-acid-CoA ligase, putative	Chlorobium tepidum	4.00E-28
9	47E10	866	NP_446059.1	fatty acid Coenzyme A ligase, long chain 5; long-chain fatty acid	Rattus norvegicus	2.00E-13
10	57A3	880	ZP_00187590.1	COG1022: Long-chain acyl-CoA synthetases (AMP-forming)	Rubrobacter xylanophilus	1.00E-27
11	57E5	378	NP_593995.1	Drug-efflux pump involved in resistance to multiple drugs; putative	S. cerevisiae	1.3
12	72E5	1345	NP_841589.1	AMP-dependent synthetase and ligase	Nitrosomonas europaea	1.00E-23
13	76B6	862	EAE96624.1	unknown	environmental sequence	5.00E-24

Figure 3-41: Contig 6 Thalassiosira pseudonana Hits

Score represented by color as follows: < 40 | 40-50 | 50-80 | 80-200 | >=200



Contig 7, Aldo/keto reductase family: Contig 7 includes seven cDNAs, which align to form a consensus sequence 1742 bp long (Figure 3-42). Overall average expression ratios were 3.12 (± 0.60) in Experiment 125C, 2.88 (± 0.26) in Experiment 125D, and 1.83 (± 0.18) in Experiment AX1 (Table 3-17). The predicted coding region for Contig 7 revealed an open reading frame that was split between frames -2 and -3 (Figure 3-43). Blast analysis indicated that both reading frames were homologous to an aldo/keto reductase conserved domain, corroborating that the deduced protein was split between two reading frames. The predicted coding regions were spliced together for subsequent analysis. Further analysis confirmed the identity of Contig 7 as likely to encode an aldo/keto reductase (Figure 3-44, Table 3-18, 3-19). Contig 7 aligned with a family of proteins that includes a number of K+ ion channels with reported oxidoreductase activity, which hints that the deduced protein may have a role in transport across the cell membrane (Figure 3-45). Alignment of contig 7 with *T. pseudonana* genome sequence revealed a region of similarity spanning over an approximately 930 bp area that appears to be interspersed with introns (overall E-value = 1.7E-35).

Figure 3-42: Sequence Alignment Overview for Contig 7 (1742bp, 7 clones, 8 sequences):

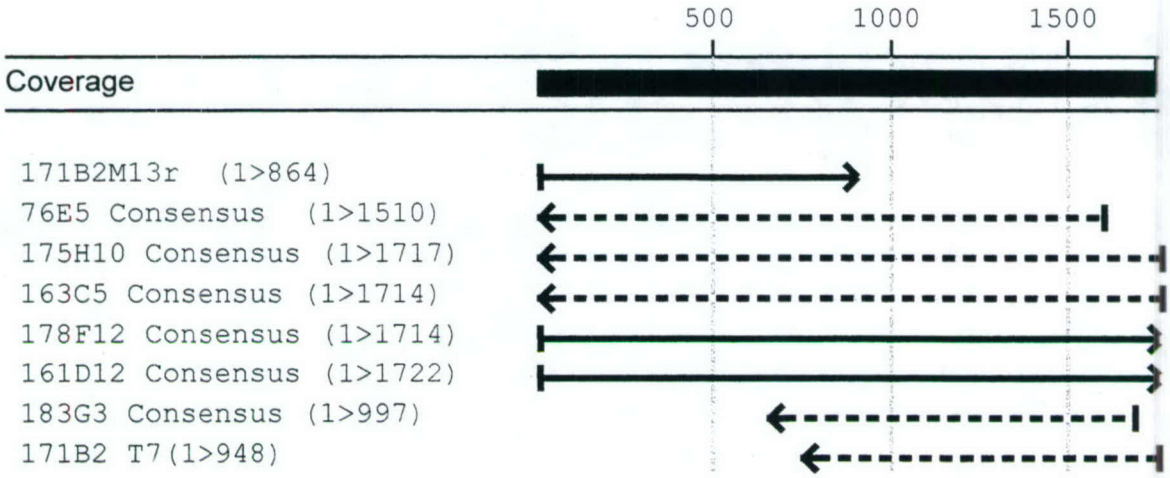


Figure 3-43: Predicted coding region for Contig 4 (ORF split between two reading frames):

Length: **150 aa**

Frame	from	to	Length
-3	335..	787	453

MGLLTGKYSNELLNENDSSPTNLSLSLKKKSNLELADLERYAYGDGVTIPEGGIQPLLKTMATIAK
RRSKTLAQVSLNYIISKGAIPIPGCRTVQQLEDNIGAMGWRLSPTEIKMLELEADKLGFGFDGAGFK
RTNEKFVGYGVEKWELD

Length: **250 aa**

Frame	from	to	Length
-2	858..	1610	753

MAFSSLVPAVAGVYILIASHFSEAYVGFSKIKPKAPIRKQFSKPLDTTKIGTLTVPTVGAGTISWSSD
NMFSTENKELEEVVREAYTSNAAFFDTAERYGSHFKTAFGMGYGETERMTSKYLVKAEQIEGPNL
LKPVVATKFTPVPWRTTVQSVVDACEQSCKNLGVEQLDLYQIHMPDIVQPLRAFGKVETKDSIYW
EGLAECYNRGLVKNVGVCNYGPTLVEQCHDALARKGVPLASNQISFSLIGET

Table 3-17: Fold-change Measurements for Individual cDNAs within Contig 7:

cDNA Identifier	125D replicate spot 1	125D replicate spot 2	125D Average	S.D.	125C replicate spot 1	125C replicate spot 2	125C Average	S.D.	AX1 replicate spot 1	AX1 replicate spot 2	AX1 Average	S.D.
161D12	2.81	**	2.81	----	2.61	2.61	2.61	0.00	1.51	1.66	1.59	0.11
163C5	3.17	3.02	3.09	0.10	2.96	2.43	2.70	0.38	1.82	1.80	1.81	0.01
171B2	3.77	4.12	3.94	0.24	3.18	3.27	3.22	0.07	2.08	2.02	2.05	0.04
175H10	3.07	3.60	3.34	0.38	3.29	3.22	3.26	0.05	2.07	2.05	2.06	0.02
178F12	3.21	3.17	3.19	0.03	2.66	2.65	2.65	0.01	1.75	1.89	1.82	0.10
183G3	1.48	2.56	2.02	0.76	2.84	2.91	2.87	0.05	1.82	1.52	1.67	0.22
76E5	3.25	3.59	3.42	0.24	3.10	2.59	2.85	0.36	1.78	1.84	1.81	0.04
AVERAGE	2.97	3.34	**3.12**		2.95	2.81	**2.88**		1.83	1.83	**1.83**	
STDEV	0.72	0.54	**0.60**		0.26	0.33	**0.26**		0.20	0.19	**0.18**	

Figure 3-44: Schematic Overview showing Contig 7 Sequence Alignments with Conserved Domain for Aldo/keto reductase:

Contig 7 Sequence Alignment with Conserved Domain for Aldo/keto reductase: Frames -2 and -3:

Contig 7 Sequence Alignment with Conserved Domain for Aldo/keto reductase: Contig 7, spliced ORF:

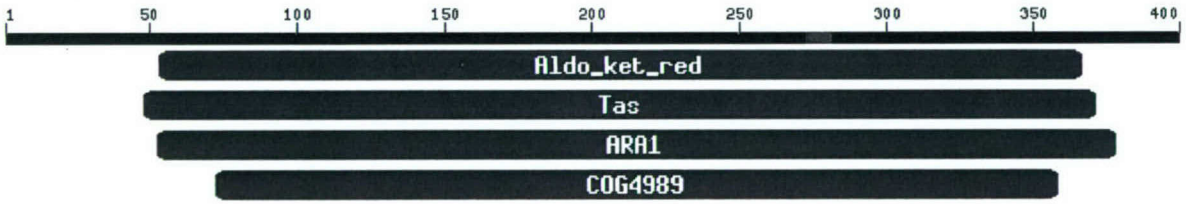


Table 3-18: Contig 7 Blast Results, Overview:

PSN Identifier	NCBI Database	Consensus Length (bp)	NCBI Identifier	Description	Species	E-value
Contig 7	CDD	1742	gnl|CDD|10536	Alcohol Oxidoreductase	COG0667	5.00E-31
Contig 7	NR	1742	NP_200170.2	aldo/keto reductase family protein	Arabidopsis thaliana	3.00E-47

Table 3-19: Contig 7, Individual cDNA BlastX Results - NR, Overview:

	PSN Identifier	Length (bp)	NCBI Identifier	Description	Species	E-value
1	161D12	1747	NP_922346.1	aldo/keto reductase	Oryza sativa	1.00E-51
2	163C5	1720	NP_200170.2	aldo/keto reductase family	Arabidopsis thaliana	2.00E-23
3	171B2	(M13r = 907, T7 = 948)	NP_922346.1	aldo/keto reductase	Oryza sativa	1.00E-14
4	175H10	1762	NP_200170.2	aldo/keto reductase family	Arabidopsis thaliana	2.00E-43
5	178F12	1722	NP_200170.2	aldo/keto reductase family	Arabidopsis thaliana	2e-26
6	183G3	997	NP_200170.2	aldo/keto reductase family	Arabidopsis thaliana	5.00E-27
7	76E5	1560	NP_200170.2	aldo/keto reductase family	Arabidopsis thaliana	4.00E-37

Contig 7, Thalassiosira pseudonana Hit:

Score represented by color as follows: < 40 | 40-50 | 50-80 | 80-200 | >=200

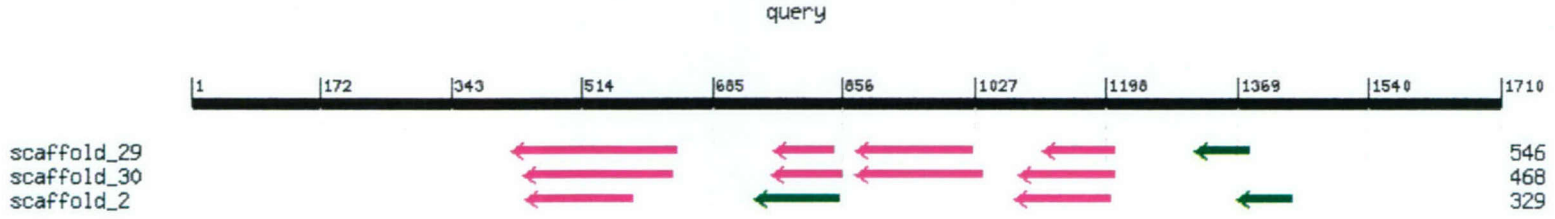


Figure 3-45: Contig 7 Sequence Aligned with pfam00248:

Aldo/keto reductase family that includes a number of K+ ion channel beta chain regulatory domains.

```
CD-Length = 279 residues,  97.1% aligned
Score =  128 bits (323), Expect = 1e-30

Query:  53   LTVPTVGAGTISWSSDNMFSTENKELEEVVREAYTSNAAFFDTAERYGSHFKTAFGMGYG  112
Sbjct:  5    LKMPLLGLGTWKTPG----QVDDEEAFEAVKAALDAGYRHFDTAEIYGN--EEEVGEAIK  58

Query:  113  ETERMTSKYLVKAEQIEGPNLLKPVVATKFTPVPWRTTVQSVVDACEQSCKNLGVEQLDL  172
Sbjct:  59   E--ALFEGSGVVREDI--------FITSKLWNT--EHSPKHVREALEKSLKRLGLDYVDL  106

Query:  173  YQIHMPDIVQPLRAFGKVETKDSIYWEGLAECYNRGLVKNVGVCNYGPTLVEQCHDALAR  232
Sbjct:  107  YLIHWPD---PLKPGDDVPIEET--WKALEKLVDEGKVRSIGVSNFSAEQLERALSEA--  159

Query:  233  KGVPLASNQISFSLIGETMGLLTGKYSNELLNENDSSPTNLSLSLKKKSNLELADLERYA  292
Sbjct:  160  RKIPPVVNQVEYHPYLRQDELR------KFCKKHGIGVTAYSPLGSGLDKGLSELGSPE-  212

Query:  293  YGDGVTIPEGGIQPLLKTMATIAKRRSKTLAQVSLNYIISKGAIPIPGCRTVQQLEDNIG  352
Sbjct:  213  ------------LLEDPALKKIAEKYGKTPAQVALRWVLQRGVSVIPKSSTPERIKENLK  260

Query:  353  AMGWRLSPTEIKMLE  367
Sbjct:  261  AFDFELTEEDMKEID  275
```

Contig 8, Unidentified: Contig 8 includes four cDNAs, which align to form a consensus sequence 1055 bp long with an open reading frame of 276aa (Figure 3-46, 3-47). Overall average expression ratios were 6.90 (± 0.93) in Experiment 125C, 5.10 (± 1.49) in Experiment 125D, and 3.01 (± 0.87) in Experiment AX1 (Table 3-20). Similarity searches revealed weak similarity (46/114; 40%) with myotubularin, which displays dual tyrosine and serine phosphatase activity (Table 3-21) (Cui et al, 1998; Laporte et al, 1998). ProSite identified two regions within the Contig 8 ORF as tyrosine sulfation sites (amino acid residues 137 - 151 *gmemdqdYtrndasl*, and 212 - 226 *alcaaddYfmepnik*). Tyrosine sulfation is a post-translational modification of many secreted and membrane-bound peptides (Figure 3-49, Moore, 2003). The up-regulation of Contig 8 in correlation with increased DA production may suggest that this protein has some role in post-translational modification of a precursor molecule leading to DA, whose destiny is ultimately to be secreted into the marine environment. Searching against the *T. Pseudonana* database revealed a 139aa region of similarity with 54% positives, 36% identity (Figure 3-48).

Figure 46: Sequence Alignment Overview for Contig 8 (1055bp, 4 clone consensus sequences):

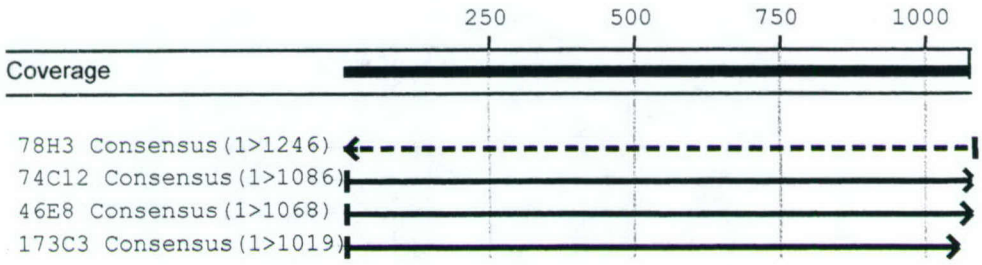


Figure 47: Predicted coding region for Contig 8:

Length: **276 aa**

Frame	from	to	Length
-1	171..	1001	831

MRSNFILLAFSLLGSSFLLDRAQGLSGLGKLVEAQYEKRVSLG
LNIGEGQNSKLAINGIVFDLMKEESRTEFSEMGKHWRASYTGGLHMLNIVQDGSFVSKQGKETVK
LKGCWEIVWREGDHGSLYCGMEMDQDYTRNDASLKGMTYVSFNAWSKEGLKKAQEFKERSAK
RANMALHKRDEELSKMLETSNIFQKGLHYYNALCAADDYFMENIKMKAVSDEEVVQFEGDMYV
SKNGKVWAHDSSKGKQVMIGTVSLELMNKQA

*Underlined regions represent ProSite identified tyrosine sulfation sites.

Table 3-20: Fold-change Measurements for Individual cDNAs within Contig 8:

cDNA Identifier	125D replicate spot 1	125D replicate spot 2	125D Average	S.D.	125C replicate spot 1	125C replicate spot 2	125C Average	S.D.	AX1 replicate spot 1	AX1 replicate spot 2	AX1 Average	S.D.
173C3	4.59	4.72	4.66	0.09	6.20	6.10	6.15	0.07	2.82	2.75	2.78	0.05
46E8	5.19	5.74	5.47	0.39	9.98	6.44	8.21	2.50	3.75	3.44	3.60	0.22
74C12	3.42	3.32	3.37	0.07	6.35	**	6.35	----	1.91	1.86	1.89	0.04
78H3	6.61	7.24	6.92	0.45	6.55	7.18	6.87	0.45	3.73	3.84	3.79	0.08
AVERAGE	4.95	5.26	**5.10**		7.27	6.58	**6.90**		3.05	2.97	**3.01**	
S.D.	1.32	1.65	**1.49**		1.81	0.55	**0.93**		0.88	0.87	**0.87**	

Table 3-21: Contig 8 Blast Results against NR database, Overview:

P. multiseries Identifier	Length	NCBI Identifier	Description	Species	Identity (%)	Positives	E-value
Contig 8	1055	NP_497766.2	myotubularin	Caenorhabditis elegans	23/114 (20%)	46/114 (40%)	6.3
78H3	1246	NP_497766.2	myotubularin	Caenorhabditis elegans	23/114 (20%)	46/114 (40%)	6.3
74C12	1086	P34756	Phosphatidylinositol-3-phosphate 5-kinase	Saccharomyces cerevisiae	11/26 (42%)	22/26 (84%)	0.94
46E8	1068	NP_738413.1	polyphosphate glucokinase	Corynebacterium efficiens	18/56 (32%)	29/56 (51%)	1.9
173C3	1019	NP_497766.2	myotubularin	Caenorhabditis elegans	23/114 (20%)	46/114 (40%)	6.3

Table 3-22: Contig 8 Alignment against Thalassiosira pseudonana:

E-value	Identities (%)	Positives (%)
1.17e-29	50/139 (35.97%)	75/139 (53.96%)

Figure 3-48: Thalassiosira pseudonana Hit

Score represented by color as follows: < 40 | 40-50 | 50-80 | 80-200 | >=200

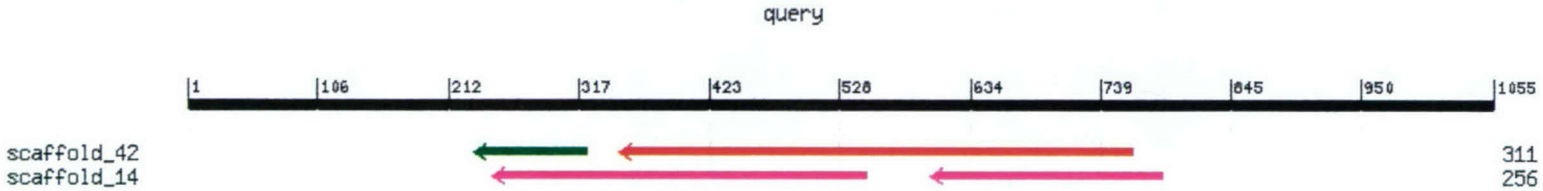
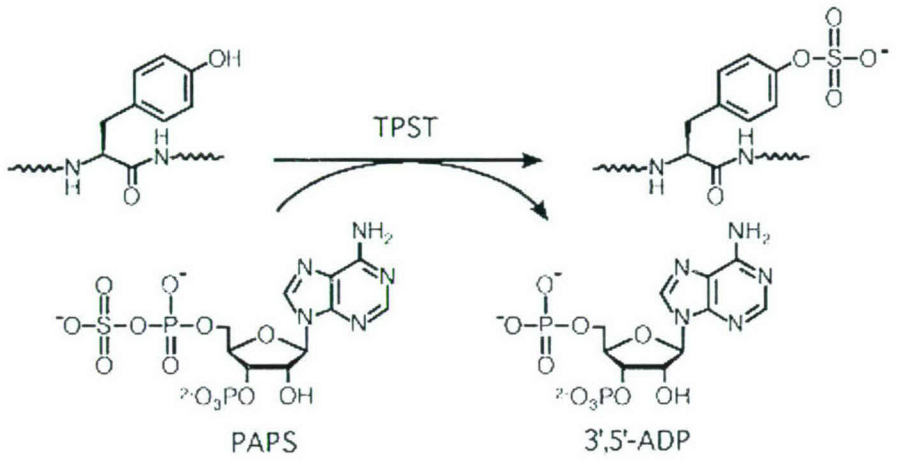


Figure 49: Tyrosine sulfation



Tyrosylprotein sulfotransferase catalyzse the transfer of sulfate from the universal sulfate donor PAPS to the hydroxyl group of a luminally oriented peptidyltyrosine residue to form a tyrosine O^4-sulfate ester and 3',5'-ADP (Moore, 2003).

Singletons: Clone 17F11 was sequenced in both directions, yielding a consensus sequence of 1270 bp with an open reading frame of 236aa (Figure 3-51, 3-52). However, gene prediction analysis of Clones 45H6, 75E8, and 6H1 revealed several possible reading frames, therefore, no single reading frame may be assigned to these cDNAs. All three of these clones were sequenced in both directions, yielding lengths of 323 bp for 45H6, 919 bp for 75E8, and two non-overlapping sequences of 552 bp and 516 bp for 6H1 (Figures 3-54 to 3-56). BLAST analysis did not reveal conclusive homologies for any of these cDNAs, however, the high fold-change values for these clones suggest that they may be promising candidates for future study (Tables 3-22, -25, -27, -29). Clones 45H6 and 75E8 demonstrated some of the highest values seen in the axenic growth experiments, and also showed high fold-change values in Experiments 125C and 125D. Homology searches did suggest that all four of these clones may encode proteins that exhibit hydrolase or isomerase activity (Tables 3-23, -26, -28, -30). For example, 6H1 appears to have some homology to a glutamine-hydrolyzing asparagine synthase with 43% identity and 53% similarity over a 30aa conserved region that is part of cd00712.1, a glutamine amidotransferase domain.

Figure 3-51:
Sequence Alignment Overview for *P. multiseries* cDNA 17 F11 (1270bp, 1 clone, 2 sequences):

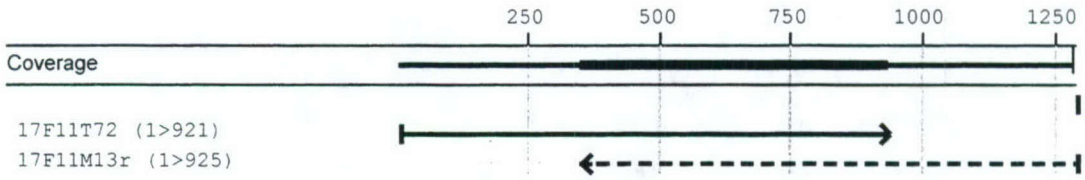


Figure 3-52: Predicted coding region for *P. multiseries* cDNA 17 F11:

Length: **236 aa**

Frame	from	to	Length
+1	49..	759	711

MDNALRDLTEHHFDPTTMSLLPSKGWESPPNYFIATTPGHPWMLMTLHYGIGSLSKIVSSMRNNP
AKHTGPSAFKIGFILFQRAIGIDTDGYLPAGIYNGAMFNGTIQQFAEAGIVLSGEAGGGKGNGTHSE
QKPQRSITLVGSKDNFHQYVDRKAIRQYSKELRKMNMSHWHEQERRPKKRVSCLEHMERQDERV
SALNLTLPSVWVPPTDMDSWWYPRYQKANYDFNGTFIEPS

Table 3-22: Fold-change Measurements for *P. multiseries* cDNA 17 F11:

	125D	125C	AX1
Replicate spot 1	5.55	6.19	2.02
Replicate spot 2	5.35	3.77	2.13
Average	**5.45**	**4.98**	**2.07**
S.D.	**0.14**	**1.71**	**0.08**

Table 3-23: *P. multiseries* cDNA 17 F11 Blast against NR database:

PSN Identifier	Length (bp)	NCBI Identifier	Description	Species	E-value	Identities (%)	Positives (%)
17F11	1270	NP_882002.1	Hydrolase (haloacid dehalogenase)	*Bordetella pertussis*	0.52	34/127 (26%)	51/127 (40%)

Talbe 3-24: *P. multiseries* cDNA 17 F11 Blast against T. pseudonana:

E-value	Identities (%)	Positives (%)
2.40E-17	36/73 (49.32%)	46/73 (63.01%)

Figure 3-53: Thalassiosira pseudonana Hit
Score represented by color as follows: < 40 | 40-50 | 50-80 | 80-200 | >=200

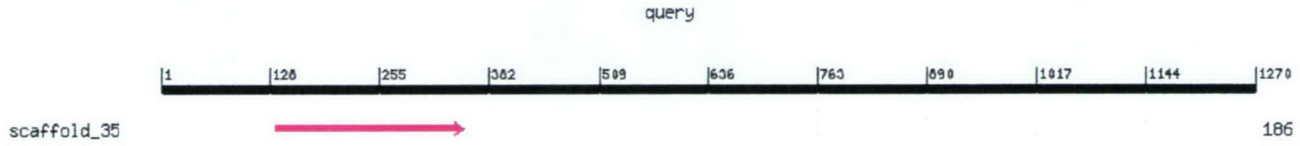


Figure 3-54:
Sequence Alignment Overview for *P. multiseries* cDNA 75E8 (919bp, 1 clone, 2 sequences):

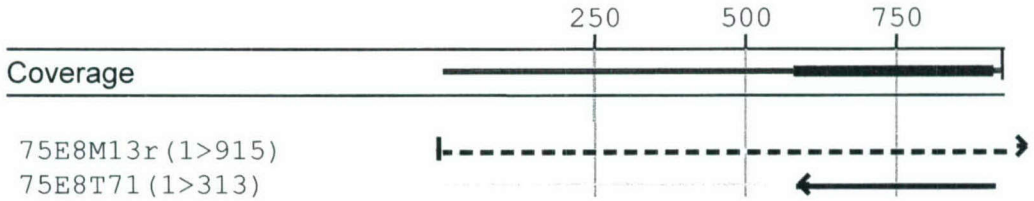


Table 3-25: Fold-change Measurements for *P. multiseries* cDNA 75E8:

	125D	125C	AX1
Replicate spot 1	3.36	3.25	3.48
Replicate spot 2	3.16	3.18	3.59
Average	**3.26**	**3.21**	**3.54**
S.D.	**0.14**	**0.05**	**0.07**

Table 3-26: *P. multiseries* cDNA 75E8 Blast against NR database:

	Consensus Length	NCBI Identifier	Description	Species	E-value	Identities (%)	Positives (%)
75E8	919	AAM54097.1|	3-O-acyltransferase	Actinosynnema pretiosum	4.7	35/122 (28%)	51/122 (41%)

Figure 3-55: Sequence Alignment Overview for *P. multiseries* cDNA 6H1 (1 clone, 2 non-overlapping sequences, M13r = 552, T7 = 516):

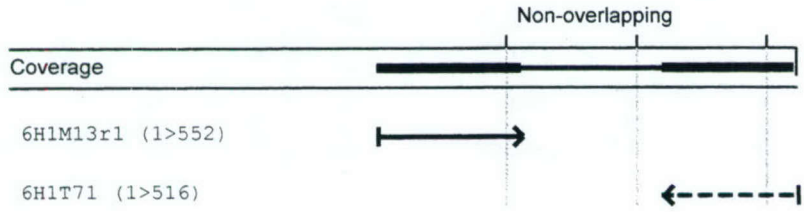


Table 3-27: Fold-change Measurements for *P. multiseries* cDNA 6H1:

	125D	125C	AX1
Replicate spot 1	5.73	2.86	2.51
Replicate spot 2	5.16	2.77	2.59
Average	**5.45**	**2.81**	**2.55**
S.D.	**0.40**	**0.06**	**0.06**

Table 3-28: *P. multiseries* cDNA 6H1 Blast against NR database:

	Length (bp)	NCBI Identifier	Description	Species	E-value	Identities (%)	Positives (%)
6H1 M13r	552	ZP_00118359.1	Asparagine synthase (glutamine -hydrolyzing)	*Cytophaga hutchinsonii*	2.0	13/30 (43%)	16/30 (53%)
6H1 T7	516	BAC55537.1	NADH dehydrogenase	*Carex shimidzensis*	1.1	20/68 (29%)	33/68 (48%)

Figure 3-56:
Sequence Alignment Overview for *P. multiseries* cDNA 45H6 (323bp, 1 clone, 3 sequences):

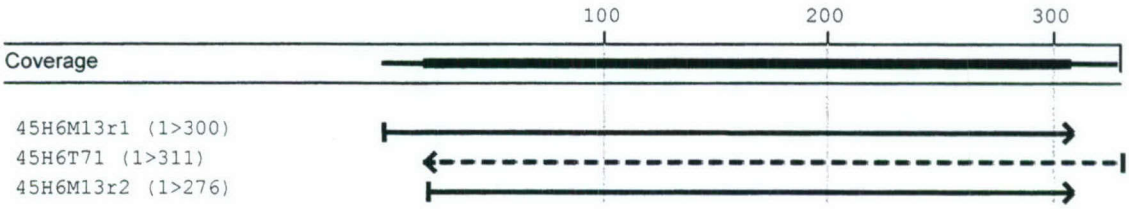


Table 3-29: Fold-change Measurements for *P. multiseries* cDNA 45H6:

	125D	125C	AX1
Replicate spot 1	3.38	3.79	3.14
Replicate spot 2	3.31	3.97	3.21
Average	**3.35**	**3.88**	**3.17**
S.D.	**0.05**	**0.13**	**0.05**

Table 3-30: *P. multiseries* cDNA 45H6 Blast against NR database:

	Consensus Length	NCBI Identifier	Description	Species	E-value	Identities (%)	Positives (%)
45H6	323	AAN39118.1|	peptidylprolyl cis-trans isomerase	Drosophila melanogaster	6.7	17/64 (26%)	28/64 (43%)

Contig 2, Novel: Contig 2 includes 53 cDNAs, which align to form a consensus sequence 2445 bp long (Figure 3-46). Overall average expression ratios were 4.25 (±0.56) in Experiment 125C, 4.24 (±0.09) in Experiment 125D, and 1.57(± 0.03) in Experiment AX1 (Table 3-31). No open reading frame could be conclusively determined for contig 2. Similarity searches revealed no significant homology with any known protein, although some cDNAs demonstrated weak similarity to glutamate-ammonia-ligase adenylyltransferase (Identities = 17/49 (34%), Positives = 25/49 (51%))(Table 3-32). No homologous sequences were found in *T. pseudonana* nor *P. tricornutum*. The consistent up-regulation among the individual cDNAs within this contig suggests that it is truly expressed, however, further investigation will be needed to determine the identity of this transcript.

Figure 3-57: Sequence Alignment Overview for Contig 2 (2445bp, 53 clones, 78sequences):

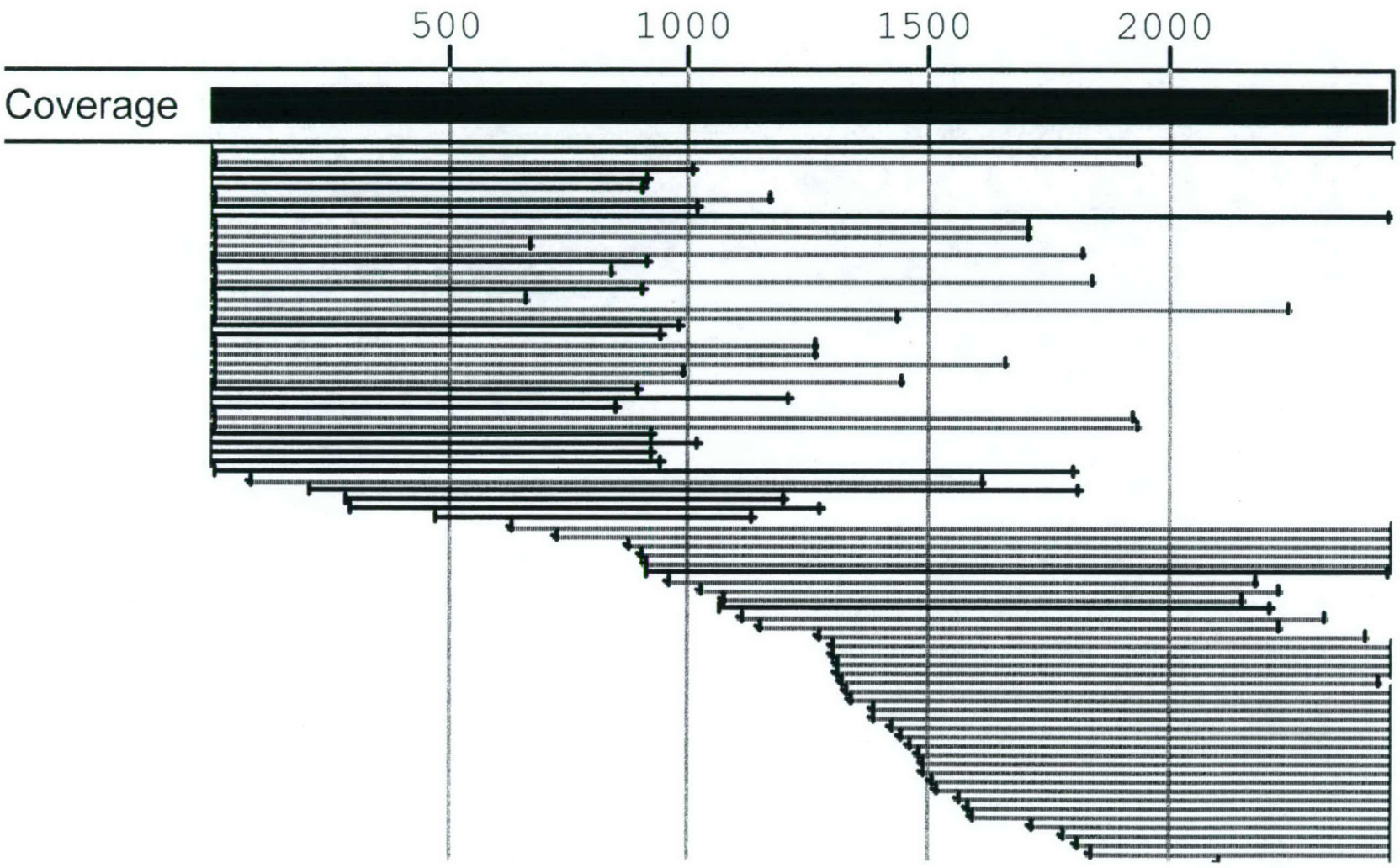


Table 3-31: Fold-change Measurements for Individual cDNAs within Contig 2:

	cDNA Identifier	125D replicate spot 1	125D replicate spot 2	125D AVG	S.D.	125C replicate spot 1	125C replicate spot 2	125C AVG	S.D.	AX1 replicate spot 1	AX1 replicate spot 2	AX1 AVG	S.D.
1	135C4	4.43	4.44	4.43	0.01	4.97	5.02	5.00	0.04	1.54	1.50	1.52	0.03
2	137A11	5.15	5.36	5.25	0.15	4.82	5.33	5.07	0.36	2.36	2.35	2.36	0.00
3	137F1	5.88	5.55	5.72	0.23	5.23	5.08	5.16	0.10	2.07	1.95	2.01	0.08
4	137F11	3.45	3.44	3.44	0.00	3.56	3.78	3.67	0.15	1.53	1.53	1.53	0.00
5	160B6	4.76	4.64	4.70	0.09	5.34	4.78	5.06	0.40	1.43	1.40	1.42	0.02
6	160H4	4.20	4.20	4.20	0.00	4.46	4.84	4.65	0.26	1.78	1.72	1.75	0.04
7	166D7	3.22	3.38	3.30	0.11	3.35	3.25	3.30	0.08	1.43	1.42	1.42	0.00
8	168D12	3.69	3.67	3.68	0.01	3.56	3.52	3.54	0.02	1.40	1.42	1.41	0.01
9	168D8	3.48	3.46	3.47	0.02	3.85	3.62	3.73	0.16	1.44	1.43	1.43	0.01
10	168F1	5.04	5.02	5.03	0.02	4.12	4.34	4.23	0.15	1.72	1.69	1.71	0.02
11	168H1	4.93	4.98	4.95	0.04	4.00	4.13	4.07	0.09	1.57	1.50	1.54	0.05
12	169D6	3.87	3.76	3.81	0.08	3.72	3.83	3.77	0.08	1.40	1.38	1.39	0.01
13	170B3	3.63	3.61	3.62	0.02	3.31	3.63	3.47	0.23	1.37	1.37	1.37	0.00
14	170C11	4.35	4.37	4.36	0.02	4.07	4.10	4.09	0.02	1.56	1.53	1.55	0.02
15	170F12	3.92	3.83	3.87	0.06	4.41	4.12	4.27	0.20	1.85	1.66	1.75	0.13
16	171B3	4.82	5.16	4.99	0.24	4.60	4.55	4.57	0.03	1.42	1.45	1.44	0.03
17	171C3	3.71	3.59	3.65	0.08	4.08	4.22	4.15	0.10	1.53	1.59	1.56	0.04
18	171F4	3.63	3.64	3.64	0.01	3.60	3.38	3.49	0.15	1.36	1.32	1.34	0.03
19	173B5	4.63	4.57	4.60	0.04	4.05	4.21	4.13	0.12	1.42	**	1.42	----
20	174C3	2.99	3.07	3.03	0.06	3.46	3.56	3.51	0.07	2.12	2.10	2.11	0.01
21	175D6	3.51	3.57	3.54	0.04	3.37	3.43	3.40	0.04	1.36	1.38	1.37	0.02
22	175H9	4.20	4.19	4.20	0.01	4.21	4.56	4.38	0.25	1.45	1.50	1.47	0.03
23	177D6	4.21	4.18	4.20	0.02	3.75	3.84	3.80	0.06	1.38	1.35	1.37	0.02
24	177E3	4.59	3.40	4.00	0.84	5.19	5.23	5.21	0.02	1.54	1.47	1.51	0.05
25	178B8	3.91	4.06	3.99	0.11	4.42	4.44	4.43	0.01	1.76	1.71	1.73	0.03
26	178B9	4.64	4.80	4.72	0.11	4.46	4.50	4.48	0.03	1.51	1.50	1.50	0.01
27	178D11	3.73	3.89	3.81	0.11	3.59	3.57	3.58	0.01	1.75	1.66	1.71	0.07
28	179C7	5.01	5.10	5.06	0.07	4.74	4.70	4.72	0.03	2.05	**	2.05	----

S.D.	cDNA Identifier	125D replicate spot 1	125D replicate spot 2	125D AVG	S.D.	125C replicate spot 1	125C replicate spot 2	125C AVG	S.D.	AX1 replicate spot 1	AX1 replicate spot 2	AX1 AVG	S.D.
29	179F10	5.11	5.31	5.21	0.15	4.79	4.69	4.74	0.07	1.64	1.62	1.63	0.01
30	179H9	4.72	4.83	4.78	0.07	4.50	4.56	4.53	0.04	1.50	1.53	1.51	0.02
31	17B7	4.86	4.75	4.80	0.08	4.74	4.61	4.67	0.10	1.53	1.57	1.55	0.03
32	17E2	3.24	**	3.24	----	4.42	4.71	4.56	0.21	1.93	1.88	1.90	0.04
33	180D12	3.64	3.63	3.63	0.01	3.47	3.74	3.60	0.19	1.49	1.51	1.50	0.01
34	186E1	3.38	3.42	3.40	0.03	4.32	4.26	4.29	0.04	1.34	1.33	1.34	0.01
35	37826F4	4.47	4.39	4.43	0.05	4.34	4.13	4.23	0.15	1.38	1.38	1.38	0.00
36	37826H7	4.92	4.95	4.94	0.02	4.02	4.23	4.13	0.15	1.44	1.45	1.44	0.01
37	45B10	4.13	4.07	4.10	0.04	3.69	3.76	3.72	0.05	1.58	1.61	1.60	0.02
38	45F9	4.00	3.86	3.93	0.10	3.67	3.66	3.67	0.01	1.69	1.67	1.68	0.01
39	47G2	4.78	4.68	4.73	0.07	4.71	4.51	4.61	0.14	1.56	1.57	1.56	0.00
40	47H11	4.54	4.49	4.52	0.03	3.81	3.74	3.77	0.05	1.61	1.56	1.58	0.04
41	51E3	4.73	4.74	4.73	0.00	4.44	4.47	4.45	0.02	**	**	----	----
42	53H4	3.83	3.93	3.88	0.07	4.43	4.30	4.37	0.09	1.59	1.64	1.62	0.04
43	54 H2 rep 1	5.19	5.25	5.22	0.04	4.22	4.27	4.24	0.04	1.39	1.43	1.41	0.02
44	54 H2 rep 2	4.57	4.51	4.54	0.05	4.53	4.83	4.68	0.21	1.47	1.43	1.45	0.02
45	55A9	3.69	3.86	3.78	0.12	4.70	4.71	4.71	0.01	1.38	1.38	1.38	0.00
46	55B4	3.67	3.67	3.67	0.00	3.70	3.60	3.65	0.07	1.48	1.46	1.47	0.01
47	5E2	3.61	3.63	3.62	0.01	4.41	4.36	4.39	0.03	1.35	1.36	1.36	0.01
48	6E1	3.38	3.42	3.40	0.03	4.47	4.34	4.41	0.09	**	**	----	----
49	72B5	5.71	**	5.71	----	4.74	5.03	4.89	0.21	1.48	1.65	1.56	0.12
50	74B5	4.94	4.92	4.93	0.02	4.45	4.65	4.55	0.14	1.37	**	1.37	----
51	75B12	3.11	3.12	3.12	0.01	3.58	3.22	3.40	0.26	1.49	1.43	1.46	0.04
52	75C4	4.26	4.64	4.45	0.27	4.59	4.49	4.54	0.06	1.75	**	1.75	----
53	77F4	3.35	3.39	3.37	0.03	3.75	3.84	3.79	0.06	1.71	1.60	1.66	0.08
54	78D8	6.12	5.25	5.68	0.61	5.33	6.31	5.82	0.69	1.87	2.01	1.94	0.10
	AVERAGE	4.25	4.22	**4.24**		4.22	4.27	**4.25**		1.58	1.56	**1.57**	
	S.D.	0.73	0.67	0.71		0.54	0.60	0.56		0.22	0.21	0.21	

Table 3-32: Contig 2 Blast Results, Overview:

	PSN Identifier	Length (bp)	NCBI Identifier	Description	Species	E-value
1	**171C3**	812	BAC87611.1	unnamed protein product	Homo sapiens	0.49
2	**186E1**	649	BAC87611.1	unnamed protein product	Homo sapiens	0.55
3	**177D6**	998	EAG90488.1	unknown	environmental sequence	0.73
4	**74B5**	822	EAA63533.1	hypothetical protein AN2962.2	Aspergillus nidulans FGSC A4	0.91
5	**17E2**	996	EAK89232.1	insulinase like peptidase	Cryptosporidium parvum	0.95
6	**53H4**	890	BAC87611.1	unnamed protein product	Homo sapiens	0.95
7	**37826F4**	906	BAC87611.1	unnamed protein product	Homo sapiens	1.1
8	**6E1**	864	BAC87611.1	unnamed protein product	Homo sapiens	1.1
9	**75B12**	950	BAC87611.1	unnamed protein product	Homo sapiens	1.1
10	**175D6**	1199	BAC87611.1	unnamed protein product	Homo sapiens	1.4
11	**55A9**	860	NP_828543.1	glutamate-ammonia-ligase adenylyltransferase	Streptomyces avermitilis MA-4680	2
12	**173B5**	820	NP_742507.1	glutamate-ammonia-ligase adenylyltransferase	Pseudomonas putida	5
13	**51E3**	952	NP_472959.2	protein kinase, putative	Plasmodium falciparum 3D7	5
14	**55B4**	1672	CAD67768.1	helical cytokine receptor CRFB7	Tetraodon nigroviridis	5.9
15	**171F4**	949	XP_345203.1	similar to hypothetical protein	Rattus norvegicus	6.4
16	**178B8**	971	NP_987388.1	tRNA pseudouridine synthase Related	Methanococcus maripaludis S2	6.6
17	**177E3**	936	AAF03787.1	Traf2 and NCK interacting kinase, splice	Homo sapiens	6.9
18	**179H9**	867	T31343	proline dehydrogenase	Bradyrhizobium japonicum	7.3
19	**45B10**	608	EAH34815.1	unknown	environmental sequence	7.4
20	**54 H2**	878	AAQ23872.1	polyprotein	Hepatitis C virus	7.4
21	**170C11**	928	AAQ23872.1	polyprotein	Hepatitis C virus	8.1
22	**168F1**	1020	T31343	proline dehydrogenase	Bradyrhizobium japonicum	8.3
23	**168H1**	976	AAQ23872.1	polyprotein	Hepatitis C virus	8.7
24	**160H4**	985	T31343	proline dehydrogenase	Bradyrhizobium japonicum	8.8
25	**37826H7**	900	AAQ23872.1	polyprotein	Hepatitis C virus	9

Table 3-32: Contig 2 Blast Results, continued:

	PSN Identifier	Length (bp)	NCBI Identifier	Description	Species	E-value
26	**175H9**	910	AAQ23872.1	polyprotein	Hepatitis C virus	9.1
27	**160B6**	914	T31343	proline dehydrogenase	Bradyrhizobium japonicum	9.2
28	**171B3**	954	AAQ23872.1	polyprotein	Hepatitis C virus	9.8
29	**17B7**	959	AAQ23872.1	polyprotein	Hepatitis C virus	9.9
30	**135C4**	946		NO HITS		
31	**137A11**	1526		NO HITS		
32	**137F1**	1623		NO HITS		
33	**137F11**	1260		NO HITS		
34	**166D7**	1194		NO HITS		
35	**168D12**	1717		NO HITS		
36	**168D8**	1601		NO HITS		
37	**169D6**	1543		NO HITS		
38	**170B3**	1297		NO HITS		
39	**170F12**	1492		NO HITS		
40	**174C3**	1150		NO HITS		
41	**178B9**	949		NO HITS		
42	**178D11**	1366		NO HITS		
43	**179C7**	733		NO HITS		
44	**179F10**	1366		NO HITS		
45	**180D12**	1591				
46	**45F9**	1484		NO HITS		
47	**47G2**	647		NO HITS		
48	**47H11**	1336		NO HITS		
49	**72B5**	1314				
50	**75C4**	1348		NO HITS		
51	**77F4**	1590		NO HITS		
52	**78D8**			NO HITS		

Down-regulated genes: Fifteen contigs were observed to be down-regulated during the transition to DA production (Table 3-33). Within this group are several genes whose likely functions can be assigned based on amino acid sequence homology. 5G12 is likely to encode *P. multiseries* ribosomal protein L22. The down-regulation of transcription of a ribosomal protein mRNA would be consistent with the switch from log phase growth to stationary phase. Similarly, sequence similarity suggests that 78B2 may represent the *P. multiseries* protein with functionality similar to the mammalian protein Kif4. This kinesin family member is a motor protein which is suggested to play an essential role in the organization of central spindles and midzone formation during cytokinesis (Kurasawa et al., 2004; Lee and Kim, 2004). Down-regulation of a gene product involved in cell division would also be consistent with the switch from log phase growth to stationary phase. The down-regulation of PSN0100, which likely encodes Ppi-phosphofructokinase is of interest because it may suggest an alteration in pathways involving energy metabolism in *P. multiseries* cells as they transition from log phase growth to stationary phase.

The down-regulation of FCP is of interest. As discussed in chapter 2, FCPs are major components of the photosystem II-associated light harvesting complex in diatoms and other brown algae (Bhaya and Grossman, 1993). Down-regulation of FCP in *P. multiseries* may be a significant aspect of the transition to stationary growth, when photosynthesis would presumably decrease as cell growth slows due to a limiting factor. The down-regulation of PSN0020, a presumptive *P. multiseries* heat shock factor 2, may represent a transition in the chaperone content of *P. multiseries* cells as they enter stationary phase.

It is also of interest to note that over half of the down-regulated genes (8 of 15) show no significant homology to any known protein coding sequence. These contigs provide an opportunity to discover new functions associated with the transition to toxin production in *P. multiseries*.

Table 3-3: Overview of Down-regulated cDNAs in PSN Differential Expression Study

PSN Library Identifier	cDNAs per Contig	Length	Putative Identification	Average Fold-change Measurements 125D	S.D.	125C	S.D.	AX1	S.D.
135H6	1	975	Fucoxanthin-chlorophyll a/c light-harvesting protein	**2.43**	0.04	**4.49**	0.33	**1.7**	0
PSN0100	2	690	PPi-phosphofructokinase	**3.01**	0.02	**3.18**	0.50	**2.69**	0.03
PSN0020	10	1149	Heat shock factor 2	**2.78**	0.41	**2.86**	0.30	**2.94**	0.68
78B2	1	867	Kif4 protein	**2.67**	0.03	**3.53**	0.11	**1.97**	0.02
5G12	1	914	ribosomal protein L22	**2.84**	**	**3.29**	0.16	**2.33**	0.01
186H3	1	760	Variable surface prolipoprotein, putative	**2.57**	0.03	**3.99**	0.08	**2.58**	0.07
PSN060	6	1975	Mucin, large thr stretch, signal peptide sequence	**5.58**	0.74	**6.40**	1.45	**2.39**	0.27
135E4	1	732	Unknown	**6.34**	0.03	**5.39**	0.54	**2.41**	0.13
136F8	1	1370	Unknown	**4.71**	0.14	**6.23**	0.16	**1.51**	0.05
137B3	1	841	Unknown	**2.70**	0.08	**2.57**	0.09	**1.53**	0.06
PSN0026	7	1373	Unknown	**2.40**	0.22	**2.84**	0.27	**6.19**	1.54
PSN0048	5	1805	Unknown	**3.05**	0.55	**4.11**	0.80	**1.96**	0.27
PSN0065	4	2056	Unknown	**3.26**	0.48	**3.33**	1.59	**1.63**	0.35
PSN0033	2	911	Unknown	**2.45**	0.08	**2.80**	0.14	**5.47**	0.14
PSN0080	3	1147	Unknown	**3.26**	0.16	**3.30**	0.23	**1.93**	0.08

**Only one value for this cDNA, replicate data insignificant.

Conclusions:

The *P. multiseries* cDNA microarray included 5372 cDNAs. Based on the redundancy calculations from Chapter 2, the number of non-redundant sequences represented on the array may be estimated as 3398, or approximately 85% of the estimated number of genes represented in the library. This suggests that an additional two or three transcriptionally up-regulated and down-regulated genes remain to be discovered within the current library, based on the parameters used in the current analysis. However, genes which are up- or down-regulated at levels below the current cutoffs will also be important to understanding the metabolic activities associated with toxin production. These transcripts remain to be discovered in the current dataset and library. It should also be noted that genes whose expression is regulated by post transcriptional mechanisms including translation will not be identified by the current microarray analysis and remain to be discovered by alternative strategies.

The analysis of *P. multiseries* transcripts following induction of DA synthesis has identified 27 transcripts of interest, twelve up-regulated and fifteen down-regulated transcripts. The further characterization of these transcripts and elucidation of their functional significance in the regulation of *P. multiseries* physiology and their potential significance in toxin production provide a series of entry points to better understand the physiology and biochemistry of *P. multiseries*. These transcripts may also be useful in ecological field studies in which they may serve as signatures of toxin production.

Chapter IV

Synthesis and Future Work

The identification and characterization of *Pseudo-nitzschia multiseries* cDNAs in this study provides an entry point for the investigation of the physiological, functional, and biochemical significance of these genes to *P. multiseries* biology and ecology. Screening the library for mRNA species which are up-regulated and down-regulated during toxin production is a first step towards fully understanding the physiological pathways that are associated with DA production in *P. multiseries*. In the immediate future, investigation of the functional role of each of these transcripts is warranted. Studies directed towards determining the causes and consequences of modulation of these genes will be of great interest. Exploration of the environmental factors that promote up-regulation and down-regulation of these genes should yield further insight into the biology of *P. multiseries*. Taken together these lines of investigation may allow the use of some or all of these transcripts as markers of *P. multiseries* physiology in the field to monitor ecologically relevant activities of *P. multiseries* such as toxin production and photosynthetic activity.

A number of immediate follow-up studies would allow more complete characterization of the transcripts identified in this thesis. The development of specific assays for each transcript of interest, through the use of quantitative PCR, RNAase protection and/or Northern blotting would be of great value. In addition to providing quantitative confirmation on the behavior of each up- or down-regulated species, these assays will allow more extensive sets of experiments to be carried out to quantitate the modulation of each mRNA species under a broad range of physiological and biochemical conditions. These detailed studies may allow the identification of mRNAs which are particularly useful as early indicators of the initiation of toxin production or the switch of *P. multiseries* cells to an alternative growth state. Northern blotting experiments will have an additional value. They may identify alternative mRNA forms which may be differentially regulated due to alternative promoters, polyadenylation or splicing. Systematic PCR or RNAase protection across each transcript would complement Northern blotting studies by revealing alternative splicing patterns which involve sequences too short to be detected on Northern blots.

The isolation and characterization of the genomic DNA which corresponds to each transcript would be of interest. It should be possible to define the promoters and regulatory elements which are responsible for the transcriptional control of these sequences. Sequence analysis and functional studies could permit the identification of key regulatory sequences responsible for the expression or repression of the transcripts we have identified during toxin production.

The functional properties of the genes we have identified can be studied in a number of ways. For many genes, charcterization by expression of a full length cDNA in an expression system such as the xenopus oocyte microinjection system could reveal some important functional characteristics of the gene. This approach would be particularly appropriate for genes involved in signaling and membrane transport. For genes involved in biochemical pathways present in microorganisms such as yeast, expression in a cell mutant for the enzyme likely to be encoded by the cDNA may be a particularly useful strategy for characterization of the enzymatic activity of the *P. multiseries* gene product.

The ability to introduce genes into *P. multiseries* for functional studies will also be of great interest. Several strategies may be useful. The development of RNA interference (RNAi) technology in *P. multiseries* to inhibit expression of target genes, would be particularly useful in determining the functional activity of the genes of interest. The ability to express genes whose expression levels are reduced or absent during a particular stage of the *P. multiseries* life cycle would also be of interest. This might allow, for example, the identification of genes which initiate the program of expression which leads to toxin production in *P. multiseries*. DNA transformation of genes into diatoms has been demonstrated using a microparticle bombardment system (Apt et al., 1996; Dunahay et al., 1995; Falciatore et al., 1999). Adaptation of a gene transfer protocol of this type to *P. multiseries* would permit experiments of the type described above to be performed.

The *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* databases have proved to be extremely useful tools in the characterization of the *Pseudo-nitzschia*

multiseries genes we have identified. Many of the *P. multiseries* sequences that were putatively identified based on sequence similarity with a known protein also matched sequences in *T. pseudonana* and *P. tricornutum*. Most often, these sequences showed the highest degree of similarity to *T pseudonana* and *P. tricornutum* sequences compared to sequences from non-diatoms. We have also identified many genes which have no significant sequence relationship to any gene in the public databases. A subset of these genes show significant sequence relationship to a gene in either the *T. pseudonana* and *P. tricornutum* databases or both. These genes should be considered to be diatom-specific transcripts. Characterization of the functional properties of these transcripts should illuminate some of the biological properties specific to the diatom family as well as the evolutionary history of diatoms.

The identification of numerous transcripts that did not match any known proteins in the public databases, nor any entry in the *T. pseudonana* and *P. tricornutum* databases may represent novel sequences that will help to elucidate unique aspects of *P. multiseries* biology, such as toxin production. The inactivation by siRNA or other methods of these transcripts in *P. multiseries* may illuminate their potentially unique role in the biology of *P. multiseries*.

Our findings have potential significance in the understanding of photosynthesis in *P. multiseries*. As noted in chapters 2 and 3, high sequence identity to known proteins substantiates the identification of Contig 3, PSN0016 as phosphoenolpyruvate carboxykinase (PCK) in *P. multiseries*. Up-regulation of this transcript was noteworthy in light of the current debate about C4 photosynthesis in diatoms. In addition, the potential identification of a C4-specific pyruvate, orthophosphate dikinase (PPDK) suggests the possibility of a C4 pathway in *P. multiseries*. C4 photosynthesis is thought to have evolved in certain plants as an adaptation to the competition of oxygen with carbon dioxide for ribulose-1,5-bisphosphate (rubisco), a key enzyme in photosynthesis. This competition occurs both during periods of high productivity, when carbon dioxide is consumed in the fixation reactions, altering the carbon dioxide to oxygen ratios in the space around the cell, and at high temperatures, when the affinity of rubisco for CO_2

decreases. Condensation of O_2 with rubisco results in the absence of fixation of CO_2. Therefore, some plants have evolved a coping mechanism in which carbon fixation involves multiple steps (Leegood, 2002). In one pattern of C4 photosynthesis, bicarbonate is fixed to oxaloacetate and transported to a separate compartment (oxaloacetate may be reduced to malate or converted to aspartate for transfer, and is then converted back into oxaloacetate), where the molecule is oxidized and decarboxylated to yield pyruvate and CO_2 by the action of PCK. The carbon dioxide may now be fixed by rubisco and photosynthesis proceeds, as in traditional C3 photosynthesis.

The diatom debate asks the question of whether C4 photosynthesis exists in diatoms. Reinfelder et al. (2000) suggest that C4 photosynthesis does exist in diatoms, in their study of *Thalassiosira weissflogii,* based on carbon labeling studies that show increased phosphoenolpyruvate carboxylase activity (which functions in carbon acquisition in C4 photosynthesis) and increased carbon pools in the form of malate during low carbon dioxide or Zn-stressed conditions. Johnston et al. (2001) dispute the conclusions made in the previous study, suggesting that further evidence is needed, including a better understanding of the role of PCK, before C4 photosynthesis can be assigned to diatoms. In the present study, up-regulation of PCK occurred under highly productive conditions with high cell densities; therefore, there is a real possibility that carbon:oxygen ratios were altered in these experiments. In addition, one proposed pathway for DA synthesis suggests that the precursor units to DA would be biosynthesized in separate compartments within the cell, which would fit the model of C4 photosynthesis (Douglas et al. 1992; Ramsey et al., 1998). These results suggest that PCK may play reciprocal roles in *P. multiseries*, potentially acting in C4 photosynthesis and DA synthesis by liberating carbon dioxide and pyruvate in the same reaction. A role for C4 carboxylation in DA synthesis has been suggested in the past (Bates, 1998). Further studies into the functional role of PCK up-regulation in *P. multiseries* will help to clarify the role of C4 photosynthesis in diatoms. In addition, localization studies with PCK or PPDK would help to determine if PCK or PPDK activity is confined within a specific compartment within the cell.

The identification of a potential amino acid transporter that is closely related to the neurotransmitter symporter family is intriguing. In plants, GABA neurotransmitter transporters appear to play a key role in signaling and pollen tube guidance in Arabidopsis (Palanivelu et al., 2000, 2003). These findings suggest the hypothesis that DA is itself a signaling molecule. An approach to expression analysis to determine functionality that would be especially useful for this transporter would be to inject mRNA into Xenopus oocytes, where the oocytes will direct synthesis of the protein so that transport function or ligand-binding properties can be assessed. In addition, antibodies against specific segments of the transport protein would help to determine which areas are exposed to one side or the other of the membrane.

The availability of *P. multiseries* cDNA microarray technology developed in this thesis offers the ability to continue expression studies to address other questions relating to DA synthesis and *P. multiseries* biology. For example, this study compared gene expression across axenic vs. non-axenic cultures in order to target *Pseudo-nitzschia* genes that are specifically related to DA production, and to reduce the likelihood of amplifying bacterial genes that may enhance toxin production. However, further analysis of this dataset and future experiments utilizing the knowledge that DA production is enhanced by bacteria would allow us to select for genes that are up-regulated in non-axenic vs. axenic cultures in order to help understand what role the bacteria have in enhancing DA production. Other experiments may focus on the effects of nutrient limitation, such as silicon limitation, which appears to enhance DA production. Another useful application of this technology will be to investigate gene expression in other *P. multiseries* species, including both toxic and non-toxic strains.

While the initial analysis of this dataset has successfully fulfilled the original goals of this project, the data generated from the microarray experiments will continue to be useful as they are annotated and analyzed further. For example, further annotating the down-regulated genes should help to broaden the picture and allow further hypotheses to be generated that will guide future research. The assessment of genes which are up or down-regulated at levels below the current cutoffs will also be of importance in

completing the picture of gene expression changes during toxin production. In addition, it will be interesting to search the dataset for specific genes that are of potential interest, such as glutamate dehydrogenase. Ramsey et al. (1998) found that the labeling pattern of carbon incorporation into DA was consistent with a biosynthetic pathway via alpha-ketoglutarate. Glutamate dehydrogenase catalyzes the reversible reaction between glutamate and alpha-ketoglutarate. Glutamate dehydrogenase expression data showed that it was up-regulated, but fell under the cut-off criteria utilized in the initial analysis. Up-regulation of this enzyme in correlation with DA synthesis supports Ramsey's model and further investigation into the functional role of this enzyme in DA biosynthesis should prove to be informative.

As is the nature of microarray experiments, the initial analysis of any one dataset is a first step into a set of data that will continue to offer useful information. The results reported here will help guide future experiments and continue to facilitate our understanding of the biochemical pathways in *P. multiseries* and other diatoms. In addition, this study demonstrates the potential of applying cDNA microarray technology to the identification of transcriptionally regulated genes in *P. multiseries* and other marine diatoms and offers a useful resource to the harmful algal bloom community.

Literature Cited:

Adams MD, Kelley JM, Gocayne JD, et al. 1991 Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651-6

Agarie S, Kai M, Takatsuji H, Ueno O 1997 Expression of C3 and C4 photosynthetic characteristics in the amphibious plant Eleocharis vivipara: structure and analysis of the expression of isogenes for pyruvate, orthophosphate dikinase. Plant Mol Biol 34:363-9

Ahn SJ, Costa J, Emanuel JR 1996 PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. Nucleic Acids Res 24:2623-5

Altschul SF, Madden TL, Schaffer AA, et al. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-402

Anderson DM 1994 Red tides. Scientific American 271:52-58

Andersson JO, Roger AJ 2002 A cyanobacterial gene in nonphotosynthetic protists--an early chloroplast acquisition in eukaryotes? Curr Biol 12:115-9

Apt KE, Droth-Pancic, PG, and Grossman 1996 Stable nuclear transformation of the diatom *Phaeodactylum tricornutum*. Mol. Gen. Genet. 252: 572-579.

Baldauf SL 2003 Phylogeny for the faint of heart: a tutorial. Trends Genet 19:345-51

Baldauf SL 2003 The deep roots of eukaryotes. Science 300:1703-6

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF 2000 A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972-7

Barraclough R, Ellis RJ 1979 The biosynthesis of ribulose bisphosphate carboxylase. Uncoupling of the synthesis of the large and small subunits in isolated soybean leaf cells. Eur J Biochem 94:165-77

Bates SS 1998 Ecophysiology and metabolism of ASP toxin production. In: Anderson, DM, Cembella, AD, Hallegraeff, GM (Eds.) , Physiological Ecology of Harmful Algal Blooms. Springer-Verlag, Heidelberg, pp. 405-426.

Bates SS 2000 Domoic-acid-producing diatoms: another genus added. J. Phycol. 36:978-983

Bates SS 2004 Interaction between bacteria and the domoic-acid-producing diatom Pseudo-nitzschia multiseries (Halse) Halse; can bacteria produce DA autonomously? Harmful Algae 3.

Bates SS, Bird CJ, Freitas ASWd, et al. 1989 Pennate diatom Nitzschia pungens as the primary source of DA, a toxin in shellfish from eastern Prince Edward Island, Canada. Can. J. Fish. Aquat. Sci. 46:1203-1215

Bates SS, Douglas DJ, Doucette GJ, LŽger C 1995 Enhancement of DA production by reintroducing bacteria to axenic cultures of the diatom Pseudo-nitzschia multiseries. Natural Toxins 3:429-435

Bates SS, Freitas ASWd, Milley JE, et al. 1991 Controls on DA production by the diatom Nitzschia pungens f. multiseries in culture: nutrients and irradiance. Can. J. Fish. Aquat. Sci. 48:1136-1144

Bates SS, Garrison DL, 1998 RAH 1998 Bloom dynamics and physiology of domoic-acid-producing Pseudo-nitzschia species. In: Anderson, DM, Cembella, AD, Hallegraeff, GM (Eds.) , Physiological Ecology of Harmful Algal Blooms. Springer-Verlag, Heidelberg, pp.267-292.

Bates SS, Hiltz MF, LŽger C 1999 DA toxicity of large new cells of Pseudo-nitzschia multiseries resulting from sexual reproduction. Can. Tech. Rep. Fish. Aquat. Sci.

Bates SS, Leger C, Smith KM 1996 DA production by the diatom Pseudo- nitzschia multiseries as a function of division rate in silicate-limited chemostat culture. In: Yasumoto T, Oshima Y, Fukuyo Y (eds) Harmful and Toxic Algal Blooms. IOC, Paris, pp 163-166

Bates SS, Richard J 1996 DA production and cell division by P. multiseries in relation to a light:dark cycle in silicate-limited chemostat culture. In: Penney R (ed) Proceedings of the Fifth Canadian Workshop on Harmful Marine Algae. Can. Tech. Rep. Fish. Aquat. Sci., pp 140-143

Bates SS, Worms J, Smith JC 1993 Effects of ammonium and nitrate on DA production by Pseudonitzschia pungens in batch culture. Can. J. Fish. Aquat. Sci. 50:1248-1254

Berman FW, LePage KT, Murray TF 2002 DA neurotoxicity in cultured cerebellar granule neurons is controlled preferentially by the NMDAreceptor Ca^{2+} influx pathway. Brain Research 924: 20-29

Bhattacharya D, Stickel SK 1994 Sequence analysis of duplicated actin genes in Lagenidium giganteum and Pythium irregulare (Oomycota). J Mol Evol 39:56-61

Bhattacharya D, Yoon H, Hackett J 2003 Photosynthetic eukaryotes unite: endosymbiosis connects the dots. BioEssays 26:50-60

Bhaya D, Grossman AR 1993 Characterization of gene clusters encoding the fucoxanthin chlorophyll proteins of the diatom Phaeodactylum tricornutum. Nucleic Acids Res 21:4458-66

Black PN, Kianian SF, DiRusso CC, Nunn WD 1985 Long-chain fatty acid transport in Escherichia coli. Cloning, mapping, and expression of the fadL gene. J Biol Chem 260:1780-9

Brown and Botstein 1999 Exploring the new world of the genome with DNA microarrays. Nature genetics supplement 21: 33-37

Botstein Ba 1999 Exploring the new world of the genome with DNA microarrays. Nature Genetics Supplement 21:33-37

Brezinski MA 1992 Cell-cycle effects on the kinetics of silicate acid uptake and resource competition among diatoms. J. Plankton Res 14:1411-1539

Cooke R, Raynal M, Laudie M, Delseny M 1997 Identification of members of gene families in Arabidopsis thaliana by contig construction from partial cDNA sequences: 106 genes encoding 50 cytoplasmic ribosomal proteins. Plant J 11:1127-40

Crepineau F, Roscoe T, Kaas R, Kloareg B, Boyen C 2000 Characterisation of complementary DNAs from the expressed sequence tag analysis of life cycle stages of Laminaria digitata (Phaeophyceae). Plant Mol Biol 43:503-13

Cui X, De Vivo I, Slany R, Miyamoto A, Firestein R, Cleary ML 1998 Association of SET domain and myotubularin-related proteins modulates growth control. Nat Genet 18:331-7

Damste JS, Muyzer G, Abbas B, et al. 2004 The rise of the rhizosolenid diatoms. Science 304:584-7

Das M, Harvey I, Chu L, Sinha M, Pelletier J 2001 Full-length cDNAs: more than just reaching the ends. Physiol Genomics 6:57-80

Davidovich NA, Bates SS 1998 Sexual reproduction in the pennate diatoms Pseudo-nitzschia multiseries and P. pseudodelicatissima (Bacillariophyceae). J. Phycol. 34:126-137

Davis R, Weintraub, H, Lassar, AB 1987 Expression of a Single Transfected cDNA converts Fibroblasts to Myoblasts. Cell 51:987-1000

Douglas DJ, Bates SS 1992 Production of DA, a neurotoxic amino acid, by an axenic culture of the marine diatom Nitzschia pungens f. multiseries Hasle. Can. J. Fish. Aquat. Sci. 49:85-90

Dunahay, TG, Jarvis, EE, and Roessler, PG, 1995 Genetic Transformation of the diatoms *Cyclotella cryptica* and *Navicula saprophila*. J. Phycol. 31: 1004-1012.

Douglas DJ, Bates SS, Bourque LA, Selvin R 1993 DA production by axenic and non-axenic cultures of the pennate diatom Nitzschia pungens f. multiseries. Elsevier Sci. Publ. B.V., Amsterdam

Edlund MB, and Stoermer, E.F. 1997 Ecological, evolutionary, and systematic significance of diatom life histories. Journal of Phycology 33:897-918

Ehara M, Watanabe KI, Ohama T 2000 Distribution of cognates of group II introns detected in mitochondrial cox1 genes of a diatom and a haptophyte. Gene 256:157-67

Eisen MB, Spellman PT, Brown PO, Botstein D 1998 Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863-8

Erdner D, Price N, Doucett G, Luisa Peleato M, Anderson DM 1999 Characterization of ferredoxin and flavodoxin as markers of iron limitation in marine phytoplankton. Marine Ecology Progress Series 184:43-53

Evans KM, Bates, S.S., Medlin, LK, Hayes, PK 2004 Microsatellite Marker Development and Genetic Variation in the Toxic Marine Diatom Pseudo-nitzschia multiseries (Bacillariophyceae). Journal of Phycology In Press

Faergeman NJ, DiRusso CC, Elberger A, Knudsen J, Black PN 1997 Disruption of the Saccharomyces cerevisiae homologue to the murine fatty acid transport protein impairs uptake and growth on long-chain fatty acids. J Biol Chem 272:8531-8

Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C 1999 Transformation of Nonselectable Reporter Genes in Marine Diatoms. 1:239-251

Fehling J, Green, DH, Davidson, K, Bolch, CJ, Bates, SS 2004 DA Production by Pseudo-nitzschia seriata (Bacillariophyceae) in Scottish Waters. Journal of Phycology In press

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P 1998 Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281:237-40

Garrison DL, Conrad SM, Eilers PP, Waldron EM 1992 Confirmation of DA production by Pseudonitzschia australis (Bacillariophyceae) cultures. J. Phycol. 28:604-607

Gasch AP, Spellman PT, Kao CM, et al. 2000 Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11:4241-57

Graycar T, Knapp M, Ganshaw G, Dauberman J, Bott R 1999 Engineered Bacillus lentus subtilisins having altered flexibility. J Mol Biol 292:97-109

Guillard RR, Ryther JH 1962 Studies of marine planktonic diatoms. I. Cyclotella nana Hustedt, and Detonula confervacea (cleve) Gran. Can J Microbiol 8:229-39

Hampson DR, Manalo JL 1998 The activation of glutamate receptors by kainic acid and DA. Nat Toxins 6:153-8

Hasle GR 1994 Pseudo-nitzschia as a genus distinct from Nitzschia (Bacillariophyceae. J. Phycol. 30:1036-1039

Hasle GR 1995 Pseudo-nitzschia pungens and P. multiseries (Bacillariophyceae): nomenclatural history, morphology, and distribution. J. Phycol. 31:428-435

Hasle GR, Lange CB, Syvertsen EE 1996 A review of Pseudo-nitzschia, with special reference to the Skagerrak, North Atlantic, and adjacent waters. HelgolŠnder Meeresunters 50:131-175

Hedrick SM, Cohen DI, Nielsen EA, Davis MM 1984 Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. Nature 308:149-53

Hildebrand M, Dahlin K, Volcani BE 1998 Characterization of a silicon transporter gene family in Cylindrotheca fusiformis: sequences, expression analysis, and identification of homologs in other diatoms. Mol Gen Genet 260:480-6

Hofte H, Desprez T, Amselem J, et al. 1993 An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from Arabidopsis thaliana. Plant J 4:1051-61

Hwang SR, Tabita FR 1989 Cloning and expression of the chloroplast-encoded rbcL and rbcS genes from the marine diatom Cylindrotheca sp. strain N1. Plant Mol Biol 13:69-79

Hwang SR, Tabita FR 1991 Acyl carrier protein-derived sequence encoded by the chloroplast genome in the marine diatom Cylindrotheca sp. strain N1. J Biol Chem 266:13492-4

Hwang SR, Tabita FR 1991 Cotranscription, deduced primary structure, and expression of the chloroplast-encoded rbcL and rbcS genes of the marine diatom Cylindrotheca sp. strain N1. J Biol Chem 266:6271-9

Johnston AM, Raven JA, Beardall J, Leegood RC 2001 Carbon fixation. Photosynthesis in a marine diatom. Nature 412:40-1

Kooistra WH, De Stefano M, Mann DG, Medlin LK 2003 The phylogeny of the diatoms. Prog Mol Subcell Biol 33:59-97

Kurasawa Y, Earnshaw WC, Mochizuki Y, Dohmae N, Todokoro K 2004 Essential roles of KIF4 and its binding partner PRC1 in organized central spindle midzone formation. Embo J

Laporte J, Blondeau F, Buj-Bello A, et al. 1998 Characterization of the myotubularin dual specificity phosphatase gene family from yeast to human. Hum Mol Genet 7:1703-12

Lea PJ, Chen ZH, Leegood RC, Walker RP 2001 Does phosphoenolpyruvate carboxykinase have a role in both amino acid and carbohydrate metabolism? Amino Acids 20:225-41

Leblanc C, Falciatore A, Watanabe M, Bowler C 1999 Semi-quantitative RT-PCR analysis of photoregulated gene expression in marine diatoms. Plant Mol Biol 40:1031-44

Lee YM, Kim W 2004 Kinesin superfamily protein member 4 (KIF4) is localized to midzone and midbody in dividing cells. Exp Mol Med 36:93-7

Leegood RC 2002 C4 photosynthesis: principles of CO2 concentration and prospects for its introduction into C3 plants. Journal of Experimental Botony 53:581-590

Lewis NI, Bates SS, McLachlan JL, Smith JC 1993 Temperature effects on growth, DA production, and morphology of the diatom Nitzschia pungens f. multiseries. In: Shimizu TJSaY (ed) Toxic phytoplankton blooms in the sea. Elsevier Sci. Publ. B.V, Amsterdam

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg S, J Q 2000 An optimized protocol for analysis of EST sequences. Nucleic Acids Res 28:3657-3665

Liang P, Averboukh L, Pardee AB 1993 Distribution and cloning of eukaryotic mRNAs by means of differential display: refinements and optimization. Nucleic Acids Res 21:3269-75

Lin S, Carpenter EJ 1998 Identification and preliminary characterization of PCNA gene in the marine phytoplankton Dunaliella tertiolecta and Isochrysis galbana. Mol Mar Biol Biotechnol 7:62-71

Lin S, Carpenter EJ 1999 A PSTTLRE-form of cdc2-like gene in the marine microalga Dunaliella tertiolecta. Gene 239:39-48

Lin S, Magaletti E, Carpenter EJ 2000 Molecular cloning and antiserum development of cyclin box in the brown tide alga Aureococcus anophagefferens. Mar Biotechnol (NY) 2:577-86

Lundholm, N. JS, Pocklington R, Moestrup ¯ 1994 DA, the toxic amino acid responsible for amnesic shellfish poisoning, now in Pseudonitzschia seriata (Bacillariophyceae) in Europe. Phycologia 33:475-478

Lundholm N, Daugbjerg, N, and Moestrup, O. 2002 Phylogeny of the Bacillariaceae with emphasis on the genus Pseudo-nitzschia (Bacillariophyceae) based on partial LSU rDNA. European Journal of Phycology 37:115-134

Mann DG 1999 The species concept in diatoms. Phycologia 38:437-495

Mann DG, Droop SJM 1996 Biodiversity, biogeography and conservation of diatoms. Hydrobiologia 336:19-32

Martin JL, Haya K, Burridge LE, Wildish DJ 1990 Nitzschia pseudodelicatissima - a source of DA in the Bay of Fundy, eastern Canada. Mar. Ecol. Prog. Ser. 67:177-182

McKay RM, Geider R, LaRoche J 1997 Physiological and Biochemical Response of the Photosynthetic Apparatus of Two Marine Diatoms to Fe Stress. Plant Physiol 114:615-622

McKay RM, Geider RJ, LaRoche J 1997 Physiological and Biochemical Response of the Photosynthetic Apparatus of Two Marine Diatoms to Fe Stress. Plant Physiol 114:615-622

Miyamoto T, Sayed MA, Sasahara R, et al. 2002 Cloning and overexpression of Bacillus cereus penicillin-binding protein 3 gene in Escherichia coli. Biosci Biotechnol Biochem 66:44-50

Moore KL 2003 The biology and enzymology of protein tyrosine O-sulfation. J Biol Chem 278:24243-6

Naidu SL, Moose SP, AK AL-S, Raines CA, Long SP 2003 Cold tolerance of C4 photosynthesis in Miscanthus x giganteus: adaptation in amounts and sequence of C4 photosynthetic enzymes. Plant Physiol 132:1688-97

Narberhaus F 2002 Alpha-crystallin-type heat shock proteins: socializing minichaperones in the context of a multichaperone network. Microbiol Mol Biol Rev 66:64-93

Ochs MF, Godwin AK 2003 Microarrays in cancer: research and applications. Biotechniques Suppl:4-15

Oeltjen A, Marquardt J, Rhiel E 2004 Differential circadian expression of genes fcp2 and fcp6 in Cyclotella cryptica. Int Microbiol 7:127-31

Okamoto OK, Hastings JW 2003 Novel Dinoflagellate Clock-Related Genes Identified Through Microarray Analysis. J. Phycol. 39:519-526

Olney JW 1994 Excitotoxins in food. Neurotoxicology 15:535-544

Palanivelu R, Brass L, Edlund AF, Preuss D 2003 Pollen tube growth and guidance is regulated by POP2, an Arabidopsis gene that controls GABA levels. Cell 114:47-59

Palanivelu R, Preuss D 2000 Pollen tube targeting and axon guidance: parallels in tip growth mechanisms. Trends Cell Biol 10:517-24

Pan Y, Bates SS, Cembella AD 1998 Environmental stress and DA production by Pseudo-nitzschia: a physiological perspective. Nat Toxins 6:127-35

Pan Y, Rao DVS, Mann KH, Brown RG, Pocklington R 1996 Effects of silicate limitation on production of DA, a neurotoxin, by the diatom Pseudonitzschia pungens f. multiseries (Hasle). I. Batch culture studies. Mar. Ecol. Prog. Ser. 131:225-233

Park T, Yi SG, Kang SH, Lee S, Lee YS, Simon R 2003 Evaluation of normalization methods for microarray data. BMC Bioinformatics 4:33

Pocklington R 1990 Trace Determination of DA in Seawater and Phytoplankton by High-Performance Liquid Chromatography of the Fluorenylmethoxycarbonyl (FMOC) Derivative. Intern. J. Environ. Anal. Chem. 38:351-368

Pohnert G 2002 Phospholipase A2 activity triggers the wound-activated chemical defense in the diatom Thalassiosira rotula. Plant Physiol 129:103-11

Quackenbush J 2002 Microarray data normalization and transformation. Nat Genet 32 Suppl:496-501

Ramsey UP, Douglas DJ, Walter JA, Wright JLC 1998 Biosynthesis of DA by the diatom Pseudo-nitzschia multiseries. Natural Toxins 6

Reap ME 1991 Nitzschia pungens Grunow f. multiseries Hasle: growth phases and toxicity of clonal cultures isolated from Galveston Bay, Texas. Texas A&M University, p 78

Reinfelder JR, Kraepiel AM, Morel FM 2000 Unicellular C4 photosynthesis in a marine diatom. Nature 407:996-9

Rhodes L, Adamson J, Scholin C 2000 Pseudo-nitzschia multistriata (Bacillariophyceae) in New Zealand. J. Mar. Freshwater Res. 34:463-467

Ridgley EL, Xiong ZH, Kaur KJ, Ruben L 1996 Genomic organization and expression of elongation factor-1 alpha genes in Trypanosoma brucei. Mol Biochem Parasitol 79:119-23

Roberts IS 1996 The biochemistry and genetics of capsular polysaccharide production in bacteria. Annu Rev Microbiol 50:285-315

Round FE, Crawford RM, Mann DG 1990 The Diatoms: Biology and Morphology of the Genera. Cambridge University Press, New York

Rudd S 2003 Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci 8:321-9

Sagerstrom CG, Sun BI, Sive HL 1997 Subtractive cloning: past, present, and future. Annu Rev Biochem 66:751-83

Sambrook J, Russell D 2001 Preparation of Plasmid DNA by Alkaline Lysis with SDS. In: Molecular Cloning A Laboratory Manual. Cold Spring Harbor Laboratory Press, New York, pp 1.32-1.34

Scala S, Carels N, Falciatore A, Chiusano ML, Bowler C 2002 Genome properties of the diatom Phaeodactylum tricornutum. Plant Physiol 129:993-1002

Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW 1996 Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A 93:10614-9

Scholin CA, Gulland F, Doucette GJ, et al. 2000 Mortality of sea lions along the central California coast linked to a toxic diatom bloom. Nature 403:80-4

Scholin CA, Villac MC, Buck KR, et al. 1994 Ribosomal DNA sequences discriminate among toxic and non-toxic Pseudonitzschia species. Nat Toxins 2:152-65

Sellner KG, Doucette GJ, Kirkpatrick GJ 2003 Harmful algal blooms: causes, impacts and detection. J Ind Microbiol Biotechnol 30:383-406

Shalon D, Smith SJ, Brown PO 1996 A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 6:639-45

Sharma V, Hudspeth ME, Meganathan R 1996 Menaquinone (vitamin K2) biosynthesis: localization and characterization of the menE gene from Escherichia coli. Gene 168:43-8

Sierra Beltran A, Palafox-Uribe M, Grajales-Montiel J, Cruz-Villacorta A, Ochoa JL 1997 Sea bird mortality at Cabo San Lucas, Mexico: evidence that toxic diatom blooms are spreading. Toxicon 35:447-53

Siezen RJ, Leunissen JA 1997 Subtilases: the superfamily of subtilisin-like serine proteases. Protein Sci 6:501-23

Sloan JL, Mager S 1999 Cloning and functional expression of a human Na(+) and Cl(-)-dependent neutral and cationic amino acid transporter B(0+). J Biol Chem 274:23740-5

Smith JC, Cormier R, Worms J, et al. 1990 Toxic blooms of the DA containing diatom Nitzschia pungens in the Cardigan River, Prince Edward Island. Elsevier Sci. Publ. Co., Inc., New York.

Smith JG, Ladinzinsky, N., and Miller, P. 2001 Amino Acid Profiles in Species and Strains of Pseudo-nitzschia from Monterery Bay California: Insights into the Metabolic Role(s) of DA. Harmful Algal Blooms 200:324-327

Smith SM, Ellis RJ 1981 Light-stimulated accumulation of transcripts of nuclear and chloroplast genes for ribulosebisphosphate carboxylase. J Mol Appl Genet 1:127-37

Spellman PT, Sherlock G, Zhang MQ, et al. 1998 Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9:3273-97

Stewart GR, Zorumski CF, Price MT, Olney JW 1990 DA: a demntia-inducing excitotoxic food poison with kainic acid receptor specificity. Exp. Neurol. 110:127-138

Subba Rao DV, Freitas ASWd, Quilliam MA, Pocklington R, Bates SS 1990 Rates of production of DA, a neurotoxic amino acid in the pennate marine diatom Nitzschia pungens. Elsevier Sci. Publ. Co., Inc., New York

Sullivan CW, Volcani BE 1973 Role of Silicon in diatom metabolism. III. The effects of silicic acid on DNA polymerase, TMP kinase and DNA synthesis in Cylindrotheca fusiformis. Biochem Biophys Acta 308:212-229

Susko E, Roger AJ 2004 Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys. Bioinformatics 127: .

Takemoto T, Daigo K 1958 Constituents of Chondria armata. Chem Pharm Bull 6:578-580

Taroncher-Oldenburg G, Griner EM, Francis CA, Ward BB 2003 Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. Appl Environ Microbiol 69:1159-71

Todd ECD 1993 DA and Amnesic shellfish poisoning - a review. J. Food Protection 56:69-83

Treguer P, Nelson DM, Van Bennekom AJ, DeMaster D, Leynaert A, Queguiner B 1995 The Silica Balance in the World Ocean: a reestimate. Science 268:375-379

Tsoi SC, Ewart KV, Penny S, et al. 2004 Identification of Immune-Relevant Genes from Atlantic Salmon Using Suppression Subtractive Hybridization. Mar Biotechnol.

Tusher VG, Tibshirani R, Chu G 2001 Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98:5116-21

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW 1995 Serial analysis of gene expression. Science 270:484-7

Venter JC, Remington K, Heidelberg JF, et al. 2004 Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74

Villac MC 1996 Synecology of the genus Pseudo-nitzschia H. Peragallo from Monterey Bay. Texas A&M University, College Station, TX, p 258

Villac MC, Roelke DL, Chavez FP, Cifuentes LA, Fryxell GA 1993 Pseudonitzschia australis and related species from the west coast of the U.S.A.: occurrence and DA production. J. Shellfish Res. 12:457-465

Wan JS, Sharp SJ, Poirier GM, et al. 1996 Cloning differentially expressed mRNAs. Nat Biotechnol 14:1685-91

Wang X 2004 Lipid signaling. Curr Opin Plant Biol 7:329-36

Werner D 1977 The Biology of Diatoms. University of California Press

Wright JLC, Boyd RK, Freitas ASWd, et al. 1989 Identification of DA, a neuroexcitatory amino acid, in toxic mussels from eastern Prince Edward Island. Can. J. Chem. 67:481-490

Wu L, Thompson DK, Li G, Hurt RA, Tiedje JM, Zhou J 2001 Development and evaluation of functional gene arrays for detection of selected genes in the environment. Appl Environ Microbiol 67:5780-90

Yueh AY, Chung CS, Lai YK 1989 Purification and molecular properties of malate dehydrogenase from the marine diatom Nitzschia alba. Biochem J 258:221-8

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO. MIT/WHOI 2004-10	2.	3. Recipient's Accession No.

4. Title and Subtitle
Gene Discovery and Expression Profiling in the Toxin-Producing Marine Diatom, *Pseudo-nitzschia multiseries* (Hasle) Hasle

5. Report Date September 2004

6.

7. Author(s) Katie Rose Boissonneault

8. Performing Organization Rept. No.

9. Performing Organization Name and Address
MIT/WHOI Joint Program in Oceanography/Applied Ocean Science & Engineering

10. Project/Task/Work Unit No. MIT/WHOI 2004-10

11. Contract(C) or Grant(G) No.
(C)
(G)

12. Sponsoring Organization Name and Address
Woods Hole Oceanographic Institution Academic Programs Office

13. Type of Report & Period Covered Ph.D. Thesis

14.

15. Supplementary Notes
This thesis should be cited as: Katie Rose Boissonneault, 2004. Gene Discovery and Expression Profiling in the Toxin-Producing Marine Diatom, *Pseudo-nitzschia multiseries* (Hasle) Hasle. Ph.D. Thesis. MIT/WHOI, 2004-10.

16. Abstract (Limit: 200 words)

Toxic algae are a growing concern in the marine environment. The marine diatom, *Pseudo-nitzschia multiseries,* produces the neurotoxin domoic acid, which is the cause of amnesic shellfish poisoning. The focus of this thesis was the molecular characterization of *P. multiseries* with the specific goal of identifying genes which may play a significant role in toxin production.

A complementary DNA (cDNA) library and a database of expressed sequence tags (ESTs) were established for *P. multiseries*. 2552 cDNAs were sequenced, generating a set of 1955 unique contigs, of which 21% demonstrated significant similarity with known protein coding sequences. Among transcripts of interest identified by sequence similarity were fucoxanthin-chlorophyll a/c light harvesting protein, C4-specific pyruvate, orthophosphate dikinase, glutamate dehydrogenase, and 5-oxo-L-prolinase.

Genes whose expression patterns were correlated with toxin production were identified by hybridization to a microarray manufactured from 5376 cDNAs. 121 cDNAs, representing 12 unique cDNA contigs, showed significantly increased expression levels in *P. multiseries* cell populations actively producing toxin. The up-regulated transcripts included cDNAs with sequence similarity to 3-carboxymuconate cyclase, phosphoenolpyruvate carboxykinase, and an amino acid transporter. Prospects for further application of molecular genetic technology to the understanding of the physiology and ecology of *P. multiseries* is discussed.

17. Document Analysis **a. Descriptors**
Gene
Toxin
Diatom

b. Identifiers/Open-Ended Terms

c. COSATI Field/Group

18. Availability Statement
Approved for publication; distribution unlimited.

19. Security Class (This Report) UNCLASSIFIED

20. Security Class (This Page)

21. No. of Pages 180

22. Price

(See ANSI-Z39.18) *See Instructions on Reverse*

OPTIONAL FORM 272 (4-77)
(Formerly NTIS-35)
Department of Commerce